2011

Prediction of crystal packing and biological protein-protein interactions with Linear Dimensionality Reduction-SVD

Sridip Banerjee
University of Windsor

PREDICTION OF CRYSTAL PACKING AND BIOLOGICAL PROTEIN-PROTEIN INTERACTIONS WITH LINEAR DIMENSIONALITY REDUCTION-SVD

by
Sridip Banerjee

A Thesis
Submitted to the Faculty of Graduate Studies
through Computer Science
in Partial Fulfillment of the Requirements for
the Degree of Master of Science at the
University of Windsor

Windsor, Ontario, Canada
2011

PREDICTION OF CRYSTAL PACKING AND BIOLOGICAL PROTEIN-PROTEIN INTERACTIONS WITH LINEAR DIMENSIONALITY REDUCTION-SVD

by

Sridip Banerjee

APPROVED BY:

Dr. Siyaram Pandey
Chemistry and Biochemistry

Dr. Alioune Ngom
Computer Science

Dr. Luis Rueda, Advisor
Computer Science

Dr. Jiangou Lu, Chair of Defense
Computer Science

8th August, 2011

Author's Declaration of Originality

I hereby certify that I am the sole author of this thesis and that no part of this thesis has been published or submitted for publication.

I certify that, to the best of my knowledge, my thesis does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my thesis, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. Furthermore, to the extent that I have included copyrighted material that surpasses the bounds of fair dealing within the meaning of the Canada Copyright Act, I certify that I have obtained a written permission from the copyright owner(s) to include such material(s) in my thesis and have included copies of such copyright clearances to my appendix.

I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee and the Graduate Studies office, and that this thesis has not been submitted for a higher degree to any other University or Institution.

Abstract

Prediction and discrimination of Crystal Packing interactions and Biological interactions is a particular problem that has drawn the attention of the research community in recent years. In this thesis, we have studied the prediction problem of these two types of interactions as well as obligate and nonobligate interactions. We are proposing new features such as Number Based Amino Acid Composition and Area Based Amino Acid Composition to predict different types of interactions more efficiently. We have measured our newly proposed features contribution to the classification by comparing them with already proposed model. Along with we are also proposing an efficient multi-stage classification strategy to successfully predict crystal packing, non-obligate and obligate interactions. In this thesis we are also proposing a modified singularity problem free linear dimensionality reductions linear transformation matrix maximization criterion. We have also applied our proposed LDR-Singular Value Decompositions modified (LDR-SVD) to other protein-protein interaction problems.

Dedication

To my parents, for all their unconditional love, caring and sacrifices they have made.

Acknowledgements

I would like to take this opportunity to express my sincere gratitude to Dr. Luis Rueda, my supervisor, for his steady encouragement, patient guidance and enlightening discussions throughout my graduate studies. Without his help, the work presented here could not have been possible.

I also wish to express my appreciation to Dr. Alioune Ngom, School of Computer Science and Dr. Siyaram Pandey, Department of Chemistry and Biochemistry for being in the committee and spending their valuable time and Dr. Jiangou Lu, School of Computer Science for serving as the chair of the defense committee.

I would like to thank to Dr. Hongbo Zhu of Max-Planck Institute of Germany who have provided me with Crystal Packing complex PDB files. Finally, I would also like to thank my friends Mominul Aziz, Mina Maleki and our Pattern Recognition and Bioinformatics lab members for their consistent moral support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Molecular biology

Molecular biology is the branch of biology that deals with molecular level details of various biological functions. In this field of study we can see overlaps with other areas of biology and chemistry, especially genetics and biochemistry. One of the main concerns of molecular biology is to understand the interactions between the various systems of a cell that includes the interactions between the different types of DNA, RNA and Protein complexes and also to figure out how these interactions are regulated. The study of molecular underpinnings of the processes of replication, transcription, translation, and cell function is summarized as molecular biology. Molecular biology devotes itself to the study of the molecular principles of the physiological processes in the life cycle of different organisms. To understand the interactions and their regulators molecular biology investigates the structure and function of various biological complexes. Molecular biology began its journey in the 1930s with the convergence of previously distinct biological disciplines: biochemistry, genetics, microbiology and virology. With the help of numerous physicists, chemists and computer scientists

molecular biology attempts to explain the phenomena of life starting from the macromolecular properties that generate them. Particularly two categories of macromolecules are the focus of the molecular biologist, nucleic acids the constituent of genes and proteins which are active agents of the living organisms.

1.2 Bioinformatics

Most of the work in molecular biology is quantitative, and those quantitative works that has been done at the interface of molecular biology, statistics and computer science is called bioinformatics. The main goal of bioinformatics is to increase the understanding of biological processes by developing and applying various computationally intensive techniques such as pattern recognition, machine learning algorithms, data mining and visualization. In this field the major research arenas are protein structure alignment, protein structure prediction, prediction of gene expression and proteinprotein interactions, sequence alignment, gene finding, genome assembly, drug design, drug discovery, and the modeling of evolution. Over the last few decades the rapid improvements in genomic and other research technologies had produced a tremendous amount of information related to molecular biology. Bioinformatics now also includes the creation and advancement of databases, algorithms, computational and statistical techniques to solve the problems arising from the management and analysis of these huge amounts of biological information. Mapping and analyzing DNA and protein sequences, aligning different DNA and protein sequences to compare them, creating and viewing 3-D models of protein structure are common activities in the field of bioinformatics. Bioinformatics was first applied to store nucleotide and amino acid sequences in a database at the beginning of ”genomic revolution”. Development of this type of databases produced not only design issues but also the development of

complex interfaces whereby researchers could access existing information as well as submit new or revised information. To understand different diseases better there is a need to study how normal cellular activities are altered in different diseased states; the biological data must be combined to form a comprehensive picture of these activities. Therefore, in the field of bioinformatics the most important task now involves the analysis and interpretation of various types of data, including nucleotide and amino acid sequences, protein domains and protein structures.

1.3 Protein-protein interaction

Protein-protein interaction prediction is an interdisciplinary field between bioinformatics and structural biology. The main goal of protein-protein interaction is to identify and catalog physical interactions between pairs or group of proteins. Prediction of different protein-protein interactions and understanding them are very important for the investigation of intercellular signalling pathways, modeling of protein complex structures and for gaining insights into various biochemical processes. Also the structural models of the complexes resulting from these interactions are necessary to understand those processes at the molecular level. Prediction of protein-protein interaction has two components, experimental approach and computational approach. Currently the work to experimentally determine the interactome of numerous species are ongoing and also computer scientists are working on developing computational methods to predict different types of interactions as accurately as possible.

In the field of proteomics one of the current goal is to map the protein interaction networks into different organisms [1]. In the complex web of interacting proteins to define a protein by its position needs protein-protein interaction information. Knowledge of this

information greatly helps biological research and makes the discovery of novel drug targets much easier. The detection of protein-protein interactions was previously limited to labor-intensive experimental techniques such as co-immunoprecipitation or affinity chromatography. Mass spectroscopy and yeast two-hybrid methods are the high-throughput experimental techniques those are now available for large-scale detection of protein-protein interactions. But these methods may not be generally applicable to all proteins in all organisms, and may also be prone to systematic error. Recently for the large-scale prediction of protein-protein interactions various complementary computational approaches have been developed based on protein sequences, structure and evolutionary relationships in complete genomes.

To understand different responsible factors for different types of interactions within protein at different level such as atomic level [11, 13, 27] different studies have been done. Also all these protein-protein interactions are not biologically relevant, there are crystal packing interactions [10] those have no biological functions associated with it. Thus there is a need to distinguish these types of interactions with biologically relevant interactions, in this thesis we are proposing a computational approach to differentiate these two types of interactions. There are diversities of protein-protein interactions[22], different types of protein-protein interactions happen in different biological processes. Among these different types of interactions obligate and nonobligate interactions [21, 35] are worth mentioning. Protomers from obligate complexes do not exist as stable structures in vivo, whereas protomers from nonobligate complexes can stay as stable and functional units. Another types of interaction those are of particular interest of research community are permanent and transient interaction, these interactions are distinguished on the basis of their lifetime. Distinguishing these types of interactions must need computational approaches because ex-

perimentally it is very hard to distinguish and study transient and permanent interactions because of their ephemeral nature. In this thesis we are proposing a computational approach to distinguish obligate and nonobligate interactions.

1.4 Feature extraction and prediction

In pattern recognition feature extraction is a special form of dimensionality reduction methods. When the input features for a classification is too large to be processed and consists of redundant data then the input features are being transformed into a reduced dimension of features, this process of reducing large dimension feature vectors to reduced dimension feature vectors is called Feature Extraction. If the extracted features are carefully chosen it is expected that the reduced feature set will extract the most relevant information from the input feature vectors. There are many popular feature extraction algorithms available namely Linear Dimensionality Reduction, Principal Component Analysis, Nonlinear Dimensionality Reduction and Independent Component Analysis. Because of its linear time complexity and higher efficiency we will use Linear Dimensionality Reduction (LDR) in our current study to predict different protein-protein interactions. In LDR we find reduced feature vectors by finding a linear transformation matrix, in this thesis we have used three different maximization criteria to get the optimum transformation matrix [26]

- Fisher's Discriminant Analysis (FDA)
- Heteroscedastic Discriminant Analysis (HDA)
- Chernoff Discriminant Analysis (CDA)

After reducing the large feature vectors to reduced feature set we have used the quadratic classifier and the linear classifier to predict different types of interactions. Quadratic clas-

sifier is a statistical classifier to separate measurements of two or more classes of objects or events by a quadratic surface. A linear classifier achieves the same classification decision based on the value of linear combinations of characteristics.

1.5 Motivation and objective

The cell is the functional basic unit of life. To perform most of the physiological functions within a cell there is a need of signal transduction. When signals from the exterior of the cell are mediated to the inside of the cell through the process of signal transduction, it happens through the chain of protein-protein interactions of signalling molecules. So we can see underneath all biological process there is some protein-protein interactions are happening. Thus it is of prime importance of research community to understand the protein-protein interactions to decipher the enigma of life. Another major importance of studying protein-protein interactions is to better understand diseases and develop drugs for them. Knowledge of these interactions information greatly helps biological research and makes the discovery of novel drug targets much easier. Because of protein-protein interactions the interacting protein complexes might be changed or it can modify other protein complexes, it means because of these interactions functionality of protein complexes might also change. Thus protein-protein interaction is also of prime importance to understand the functionality of protein complexes.

Previously there were only labor-intensive approaches such as affinity chromatography or co-immunoprecipitation was available. Currently high throughput experimental techniques such as mass spectroscopy and yeast two hybrid methods are available for large detection of protein-protein interactions. But these methods may not generally be applicable to all proteins in all organisms, and may also be prone to system error. For that reason

recently for large-scale prediction of protein-protein interactions various complementary computational approaches have been developed.

Among different types of protein-protein interactions we have focused on distinguishing biological interactions with crystal packing interactions [35]. Previously computational approaches have shown that biological interactions have larger interface size than non-biological interactions [10, 23]. We have also predicted obligate and nonobligate protein-protein interactions successfully with more classification accuracy [21, 35]. Determining nonobligate and obligate interactions with experimental approaches such as co-immunoprecipitation or affinity chromatography often leads to erratic results. Thus efficient computational approaches were necessary to successfully predict these types of interactions. We have also devised a multi-class classification model to classify biological and non-biological interactions and nonobligate and obligate interactions.

To give an idea how rapidly research is going on this particular field, we can see the growth of identifying protein three-dimensional structures by the research communities, Protein Data Bank where the three-dimensional structures of different protein complexes are kept started in 1971 with only 7 structures, in 2003 it had 20,000 structures and in 2010 it had 68,000 structures.

1.6 Problem statement

To predict different protein-protein interactions(biological-crystal packing and obligate-nonobligate)

To solve the problem we propose a two step process:

1)Generation of features from physio-chemical properties of the protein complexes

2)Prediction with linear dimensionality reduction-SVD

1.7 Contribution

The main contributions of this thesis are:

- Propose 40 new features on Solvent Accessible Surface Area (SASA) property to predict biological-crystal packing interactions and obligate-nonobligate interactions and compare with previous approaches to verify the effectiveness of our new proposed features by showing increased prediction accuracy.
- Propose a multi-class classification model to predict biological(nonobligate-obligate)-crystal packing interactions and compare them with previous approaches.
- Propose a solution to the generation of singularity matrix problem for CDA criterion of linear dimensionality reduction.
- Implement different visual classification analyzer tool such as Receiver Operating Characteristic (ROC) Curve and Matthews correlation coefficient to analyze the effectiveness of our prediction.

1.8 Thesis organization

This thesis has six chapters. A survey of crystal packing-biological interactions and nonobligate-obligate interactions is presented in Chapter II. Chapter III presents a detailed discussion about different feature extraction and pattern classification methods. In Chapter IV we describe newly proposed features, solution to the singularity problem for CDA criterion for linear dimensionality reduction and multi-class classification model. In Chapter V we have shown experimental results and findings with proposed approach and its comparison with

the existing methods. At last in Chapter VI we conclude the thesis and identify future works that can be extended from this work.

Chapter 2

Protein-protein interaction prediction

2.1 Protein

Proteins are nitrogenous organic compounds that consist of large molecules of one or more long chains of amino acids and are an essential part of all living organisms. Proteins consist of one or more polypeptides folded into a globular form in a biologically functional way. A polypeptide is a single polymer chain of amino acids bonded together by peptide bonds between the carboxyl and amino groups of adjacent amino acid residues. Proteins are structural components of body tissues such as muscle, hair, collagen, etc. and as enzymes and antibodies.

2.2 Proteomics

Proteomics is a field of study that deals with structural and functional properties of proteins. Proteins are the main components of physiological metabolic pathways of cells, so they are of prime importance of study to understand the mechanism of any living organisms.

The term was first coined in 1997, to make an analogy with 'genomics'. Proteomics is considered much more complicated than genomics because while an organism's genome is more or less constant, proteome (proteome is an entire complement of proteins, including the modifications made to a particular set of proteins produced by an organism or system) differs from time to time and from one cell to a different cell.

Proteomics gives much more better understanding of an organism than genomics. For that reason scientists are very interested in proteomics lately. First, the level of transcription of a gene gives only a rough estimate of its level of expression into a protein. It is known that mRNA is translated into protein, but it is not always true. It is possible that an mRNA produced in abundance may degrade rapidly or translate inefficiently resulting in a small amount of protein. Second, many proteins experience post-transitional modifications that profoundly affect their activities, as an example some proteins are not active until they become phosphorylated. Third, many transcripts give rise to more than one protein through alternative post-transitional modifications. Fourth, many proteins form complexes with other proteins or RNA molecules, and only function on the presence of these molecules. Finally, protein degradation plays an important role in protein content [4].

The benefits of studying human genes and proteins are the identification of the potential new drugs for the treatment of disease. Scientists at first obtain genome and proteome information that identifies proteins associated with a disease. Then after identification of that vicious protein, they look into the three-dimensional structure of that protein, which provides all the information to design drugs to interfere with the action of the protein.

Most proteins function in collaboration with other proteins. Thus, it is one of the major goal of proteomics is to identify which proteins interact with other proteins. There are several methods available to probe protein-protein interactions. The very popular traditional

method is yeast two-hybrid analysis. There are several new methods those are currently used by research community namely protein microarrays, immunoaffinity chromatography, mass spectrometry, dual polarisation interferometry, microscale thermophoresis and experimental methods such as phage display. Currently there is a substantial research effort undergoing to develop effective computational methods to predict and identify different protein-protein interactions and this is one of the major contributions of this thesis and study.

2.3 Protein structures

Proteins are biochemical compounds consisting of one or more polypeptides. A polypeptide is a single linear polymer chain of amino acids bonded by peptide bonds. To understand the functions of protein sometimes it is necessary to determine their three-dimensional structure. This topic is the field of study in structural biology, which uses different techniques namely X-ray crystallography, NMR spectroscopy, and dual polarization interferometry to determine the structure of proteins. There are four distinct levels of protein structure primary structure, secondary structure, tertiary structure, quaternary structure.

The primary structure (Figure 2.1), of a protein is expressed by the amino acid sequences of the polypeptide chain. The sequence of a protein is unique to that protein. The amino acid sequence of a protein can be determined by Edman degradation or tandem mass spectrometry methods. The primary structure (Figure 2.1) of a protein is held together by peptide bonds.

The secondary structure of a protein (Figure 2.2)describes the regions of the chains of a protein that are organized into local sub structures known as alpha-helices or beta-sheets. These secondary structures are held by hydrogen bonds between the main chain peptide

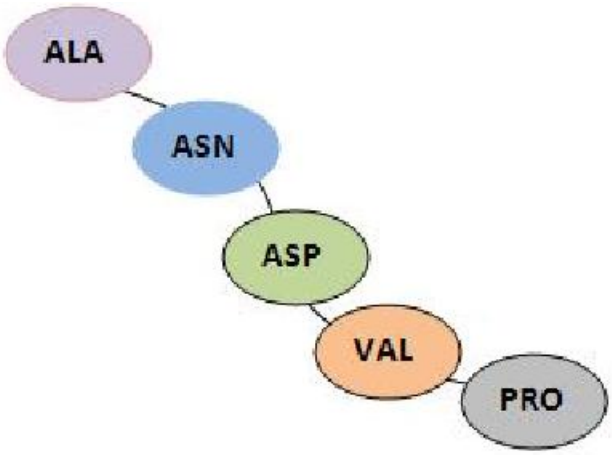

Figure 2.1: Primary Structure of 1AVA A B C protein complex. Generated with the help of ICM Molsoft browser

groups.

The tertiary structure (Figure 2.3) of protein molecules describes the specific atomic positions in three-dimensional space. In tertiary structure alpha-helices and beta-sheets are folded into a compact globule. The folding is driven by non-specific hydrophobic interactions. In this thesis and study we have worked with this structure of proteins and used three-dimensional shape information for the generation of feature vectors for the prediction.

The quaternary structure (Figure 2.4) of a protein describes the structure of a protein which has more than one polypeptide chain which are also called subunits.

Figure 2.2: Secondary Structure of 1AVA A B C protein complex. Generated with the help of ICM Molsoft browser

2.4 Protein-protein interactions

Protein-protein interaction happens when two or more proteins bind together mostly to carry out their biological functions. Protein-protein interactions are at the main part of the entire interactomics (interactions and the consequences of the interactions between and among proteins or other molecules within a cell) system of any living cell. Protein-protein interactions are important for the majority of biological functions. Most of the molecular processes happen in the cell because of protein-protein interactions, such as DNA replication carried out by large molecular machines organized by their protein-protein interactions. Enzymes interact with their substrates, inhibitors interact with enzymes, transport proteins interact with structural proteins, hormones interact with receptors and these are small subsets of interactions that happen in a cell. Signals from the exterior of the cell are mediated to the inside of the cell by protein-protein interactions of those signalling molecules. This process is known as signal transduction and it plays a crucial role in many biological pro-

Figure 2.3: Tertiary Structure of 1AVA A B C protein complex. Generated with the help of ICM Molsoft browser

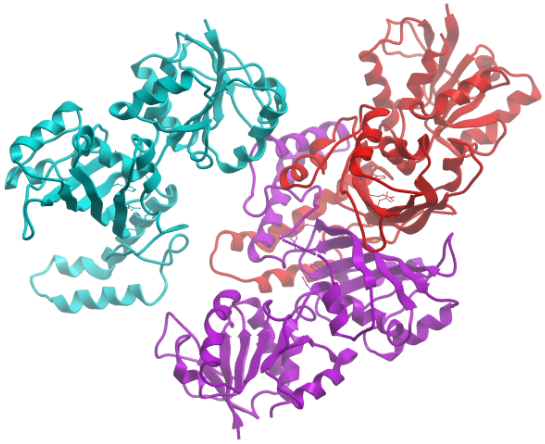

Figure 2.4: Quaternary Structure of 1AVA A B C protein complex. Generated with the help of ICM Molsoft browser

cesses and diseases such as cancer. Protein might interact for a long time to form a part of the protein complex, a protein might carry another protein such as to carry a protein from cytoplasm to nucleus or vice versa, a protein might interact with another protein to modify it. These are the reasons why proteins interact. Specifically protein-protein interactions happen because every protein wants to be more stable to have a specific shape that performs particular functions.

Molecules belong to specific amino acids within a protein, interact with each other if the distances between them are $1\mathring{A}$ - $7\mathring{A}$. These interactions between molecules are called direct contact association of molecules. Long range interactions (if the molecules are more than $7\mathring{A}$ apart) of molecules are also possible if the surrounding neighborhood such as water solution helps molecules to interact with each other. In (Figure 2.5) complex 1B3A has 2 chains and they are in direct contact with the atoms of the different chains in the marked area.

Not only the interactions but also the structural models of the complexes resulting from these protein-protein interactions are necessary to understand most of the biological processes at the molecular level. Protein-protein interactions have been studied from the perspectives of biochemistry, quantum chemistry, molecular dynamics, signal transduction and other metabolic or genetic networks. A huge active research is currently going on to probe different protein-protein interactions by applying different methods including computational methods, which is the main topic of the current study.

2.4.1 Different types of protein-protein interactions

Different types of protein-protein interactions vary on the basis of their belongings to particular protein family and their different three-dimensional structures. Protein-protein inter-

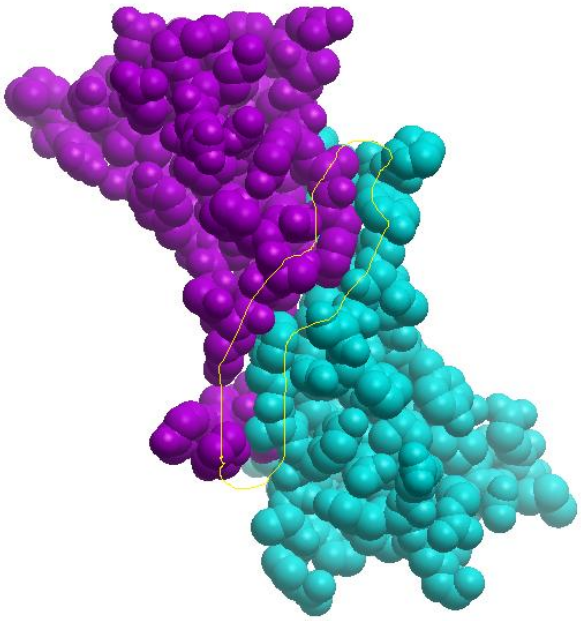

Figure 2.5: Protein-protein interaction of 1B3A Obligate Interactions between Chain A and B. Generated with the help of ICM Molsoft browser

actions play different roles in different biological processes starting from signalling carrier to another protein generation. By studying protein-protein interactions we can know what the behavior of particular interacting protein molecules is. On the basis of physiological functions, specificity and evolution, protein-protein interactions can be divided into 4 categories [22] such as Homo- and Hetero-oligomeric, Obligate and Non-obligate, Transient and Permanent, Crystal packing and Biological [35].

Homo-oligomer and hetero-oligomeric protein-protein interactions: If two interacting protein chains of an oligomer have structural symmetry then those kinds of protein-protein interactions are called homo-oligomeric protein-protein interactions. If the two interacting chains of oligomers have difference in their structure, then that kind of interaction is classified as hetero-oligomeric protein-protein interaction.

Obligate and non-obligate protein-protein interactions: If the interacting protomers are

not found as stable structures in vivo, that kind of protein-protein interaction is called obligate protein-protein interaction. If the interacting protomers have stable structure in vivo then that kind of protein-protein interaction is classified as non-obligate protein-protein interaction.

Transient and permanent protein-protein interactions: These two different types of protein-protein interactions are differentiated on the basis of the life-time of the interacting protein complexes. Permanent interactions outputs stable complexes but the transient protein-protein interactions outputs less stable complexes. These less stable complexes that are being generated from transient interactions continuously change their shapes until they convert into a stable complex.

Crystal packing and biological protein-protein interactions: Different protein complexes during crystallization process (as an example: X-ray crystallography) form solid crystals. This kind of interaction do not serve any biological purpose, because by forming crystal of different protein complexes they do not carry out any biological functions. This kind of protein-protein interaction is called crystal packing interaction. Most of the crystallized protein complexes that are being produced as a result of crystal-packing interactions does not exist in vivo. All other protein-protein interactions that happen and also carry out some biological functions are called biological protein-protein interactions. In this thesis we have predicted crystal packing and biological protein-protein interactions and also obligate and non-obligate interactions.

2.5 Crystallization of proteins

X-ray crystallography plays a central role to determine the structural models of proteins. Among all the other methods available it is the most popular method [25]. It is a method

to determine the arrangements of atoms within a crystal. In this method, a beam of X-rays strikes a crystal and diffracts it into many specific directions. By measuring the angles and intensities of these diffracted beams a crystallographer can generate a three-dimensional picture of the density of electrons within the crystal. The mean positions of the atoms in the crystal, their chemical bonds and their disorder and other different information can be determined from this electron density. The importance of protein crystallization is that it is the basis of X-ray crystallography.

Protein molecules can be crystallized if the purified protein undergoes slow precipitation from an aqueous solution. As a result of the crystallization, individual molecules align themselves in a repeating series of unit cells by adopting a consistent orientation. The crystallized protein complexes that results from the crystallization process are held together by non-covalent interactions. It is observed that protein crystals if surrounded by their mother liquor gave better diffraction patterns than dried crystals [2].

The primary goal of crystallization of protein is to produce a well-ordered crystal that lacks contaminants. The generated crystals should also be large enough to provide a diffraction pattern when it would be exposed to X-rays. The diffraction pattern that was generated by the X-rays then should be analyzed to discern the protein's tertiary structure.

2.6 Crystal packing and biological protein-protein interactions

As discussed in the previous section, X-ray crystallography is the most popular method to determine three dimensional structures of protein complexes. But not all interactions those are observed in structures of protein complexes determined by X-ray crystallography are

biologically relevant. Many of these interactions that are observed form during the crystallization process. These interactions would not appear within any living organisms. These crystal packing contacts are non-specific and there are no biological functions associated with them. Thus there is a huge need to discriminate these kinds of interactions with biologically relevant interactions, as we do not want to waste our time studying protein-protein interactions those have no biological relevance.

Previous studies have examined properties of protein-protein interfaces in order to differentiate between biologically relevant interactions and non-biological interactions those results from crystal packing contacts. It has been observed that biological interactions tend to have large protein interface sizes than non biological interactions [6, 9, 10, 15, 23, 28]. In PQS [15] they have primarily used interface size as the main discriminant to separate true biological interactions with crystal packing interactions with a 78% accuracy on a non-redundant dataset [33]. PQS used 400 $\AA^2$ cut-off for interface size to discriminate between biological interactions and non-biological interactions. Amino acid composition of the interface is also another well-analyzed discriminating feature for identifying biological interactions [12, 16, 23, 24]. It has been found that amino acid composition of biological interfaces is different from that of the rest of protein surface [12, 16, 24]. On the other side, in his paper [23] Carugo and co-authors showed that the chemical composition of interface of crystal packing contacts is very similar to that of the rest of the surface as a whole. In [35] Zhu et al. used six physio-chemical interface properties to discriminate biological interactions with crystal-packing contacts with 90.9% accuracy. These interface properties are also used to identify different protein-protein interaction sites. Jones and Thornton analyzed six physio-chemical interface properties to predict different interaction sites [12, 31].

2.7 Obligate and non-obligate protein-protein interactions

We can differentiate biological protein-protein interactions based on different properties such as homo-oligomer and hetero-oligomer, obligate and nonobligate, transient and permanent. In our study we discriminate between obligate and nonobligate protein-protein interactions. If the interacting protomers are not found as stable structures in vivo that kind of protein-protein interaction is called obligate protein-protein interaction. If the interacting protomers have stable structure in vivo then that kind of protein-protein interaction is classified as non-obligate protein-protein interaction.

There are several studies have been done previously to differentiate obligate and non-obligate protein-protein interactions. The study [8] by Nooren et al revealed that interfaces of nonobligate complexes have smaller area, and are more planar and polar on average than those of stable homodimers. It has been also found that interface residues of nonobligate homodimers be more conserved than the other surface residues. In their study [14], Gunasekaran et al reported that both pre-residue surface area and interface area of non-obligate interactions are much smaller than those of obligate interactions. In this study [29], De et al. performed a statistical analysis of the interface properties for obligate and non-obligate interactions. They reported that obligate protein-protein interaction interfaces have more contacts than non-obligate interfaces, and these contacts are nonpolar. Mintseris et al.[21] have explained the difference between obligate and nonobligate complexes from an evolutionary point of view. Interface residues were reported to be significantly more conserved in obligate interactions than those in non-obligate interactions. That study also showed that the coevolution rate is lower for obligate interactions than for non-obligate interactions. In general, obligate and non-obligate proteins have distinct interaction preferences. The point is that there is no single interface property with a clear cut-off on that basis we can

discriminate between different protein interaction types. But this is explainable given the complexity and diversity of protein interactions.

In the NOXClass study, [35] Zhu et al have investigated six interface properties namely

1. Interface area
2. Ratio of interface area to protein surface area
3. Amino acid composition of the interface
4. Correlation between amino acid compositions of interface and protein surface
5. Interface shape complimentarity
6. Conservation score of the interface

By taking into account these protein-protein interaction interface properties they predict biological and non-biological interactions and intra-biological (obligate and non-obligate) interactions. As in the study in the thesis we propose 40 new computed derived features from these interface properties and predict different interaction types with higher accuracy than the NOXClass [35], it is worth discussing here in extenso these interface properties.

Interface Area: Interface area of a protein protein interaction is defined as one half of the total decrease of SASA(Δ SASA) (Solvent Accessible Surface Area) of the two protomers upon the formation of interaction.

$$InterfaceArea = \frac{1}{2}(SASA_a + SASA_b - SASA_{ab}) \tag{2.1}$$

In the above equation a and b are two protomers in the complex ab; and $SASA_a$, $SASA_b$ and $SASA_{ab}$ are the values for a, b, and ab respectively. A residue is defined as being part of the protein-protein interaction interface if its Solvent Accessible Surface Area (SASA) decreases by greater than 1 $\AA^2$ upon the formation of the complex [30]. SASA values for residues were calculated using NACCESS [7], with a probe sphere of radius 1.4 $\AA^2$.

Interface Area Ratio: In biological protein-protein interaction involving small protomers can not have large interface areas for instance in some enzyme-inhibitor complexes. For that reason we have normalized interface area by the SASA of the smaller protomer in the complex.

$$InterfaceAreaRatio = \frac{InterfaceArea}{min(SASA_a, SASA_b)} \tag{2.2}$$

In the above equation $SASA_a$ and $SASA_b$ are the SASA values for interacting protomers a and b respectively.

Amino Acid Composition of the interface: To obtain the amino acid composition of the interface we have we have calculated both number based and area based amino acid composition [24]. The number-based amino acid composition (v_n) of the interface is calculated as the frequency of each type of the 20 standard amino acids in the protein-protein interface. By weighting each residue with its Δ SASA, the area-based amino acid composition of protein-protein interaction interface v_a is computed:

$$v_{a,i=1,\ldots,20} = \frac{1}{2\ Interface\ Area} \Sigma_{r,type(r)=i} \Delta SASA(r) \tag{2.3}$$

In the above equation type(r) is the type of the amino acid of residue r.

The amino acid composition of the interface is calculated by this equation where Δv distance between two vectors v and $v\prime$ of amino acid composition, number or area based [16, 24].

$$(\Delta v)^2 = \frac{1}{19} \Sigma_{i=1}^{20} (v_i - v_i\prime)^2 \tag{2.4}$$

Correlation between amino acid compositions of interface and protein surface: In [34] Ofran et. al. showed that the amino acid composition of the biological interface to be significantly different from that of the rest of the protein surface. We can expect that amino acid composition of the crystal packing interface to be similar to the rest of the protein surface. Thus, it can be a good distinguishable feature to discriminate between biological and non-biological interaction types. To capture this effect, the Pearson correlation coefficient between the amino acid composition of interface and surface was calculated.

Gap Volume Index: It has been shown in [12, 24] that protein-protein interfaces are more complimentary in nature in obligate protein complexes than those in non-obligate complexes. It is shown by Bahadur et al [24] that the gap volume index is one of the measurements for interface complementarity. As the gap volume is dependent on protein size,

it is computed by normalizing the gap volume between protomers with their interface area.

$$GapVolumeIndex = \frac{GapVolume}{InterfaceArea} \tag{2.5}$$

The more complementary the interface shapes are the smaller the gap volume index would be. The gap volume was computed using the SURFNET program [17]. The minimum and maximum radii for gap sphere were set to 1.0 to 5.0 $\AA$ respectively for the computation of the gap volume index, and the grid separation was set to 2.0 $\AA$.

Conservation scores of the interface: Conservation scores for residues in the interface were calculated by Consurf [3]. The average value of conservation scores of all the residues at the protein-protein interface is defined as the conservation score of the interface. As we have weighted area based amino acid composition, in a similar fashion we weighted the conservation score for each residue by its ΔSASA upon the formation of the interaction. The average of these weighted residue conservation scores was used as the area-based conservation score of the interface.

Chapter 3

Feature extraction and prediction

3.1 Pattern recognition

Pattern recognition is the scientific field whose goal is the classification of objects into a number of categories and classes [32]. These objects can be images, protein sequences, signal waveforms, email messages or any type of measurements that need to be classified depending on the application. The field of pattern recognition has a long history, but before the 1960s it was primarily the output of theoretical research in the area of statistics. With the advent of computers and digital devices there was an increase in the demand of practical applications of pattern recognition, which in turn set new demands for further theoretical developments in this field. As information retrieval and handling are becoming the most important activities in modern days, so the demands of pattern recognition techniques are also rising. These demands have pushed pattern recognition to the high edge of today's engineering application and research.

There are several application areas of pattern recognition in today's world. It is an inseparable part of most machine intelligence systems built for decision making. Pattern

recognition is also of high importance in the area of machine vision. A machine vision system takes images via a camera and analyzes them to produce descriptions of that image. An example of a machine vision system is an automated visual inspection system in the assembly line of a manufacturing firm. Pattern recognition also has applications in the field of automated information handing, as is the example of Optical Character Recognition systems. A typical commercial application of an optical character recognition system would be an automated machine that reads bank checks, this machine must be able to recognize the amounts in figures and digits and match them. Another application of optical character recognition system is an automatic mail sorting machines for postal code identifications in post offices. Pattern recognition also has applications in medical fields such as computer-aided diagnosis. It aims to assist doctors to make diagnostic decisions. The medical data are not often easily interpretable. Computer aided diagnosis plays an important role there by automatically interpreting X-rays, computer tomographic images, ultrasound images and electrocardiograms. Pattern recognition has its contributions to the field of speech recognition. Recently a huge research and development effort has been invested in the field of speech recognition. Speech is the most natural means by which human beings exchange information among each other. For that reason building intelligent machines that recognize spoken information has been of prime importance, and pattern recognition has a significant contribution to achieve that goal.

Pattern recognition algorithms are also applied to the field of fingerprint identification, signature authentication, face recognition, gesture recognition and text retrieval [32]. Face and gesture recognition have recently attracted much research interest and investment in an attempt to facilitate human-computer interaction and to further enhance the role of computers in office automation, automatic personalization of environments [32]. Pattern recogni-

tion is also closely linked to other scientific disciplines such as linguistics, computer graphics and computer vision. In this thesis we apply pattern recognition to successfully predict different types of protein-protein interactions. This is of high importance in proteomics research.

In pattern recognition to classify objects in the categories and classes at first measurable quantities of information are collected from these objects. These objects are called samples and the observed numerical properties that are collected from these samples are called features. These features are then fed to the classifiers to successfully classify or predict different samples.

3.2 Feature extraction methods

When the input features for a classification are too large to be processed and consists of redundant data, the input features are transformed into a reduced dimension of features. This process of reducing large dimension feature vectors to reduced dimension feature vectors is called feature extraction [5]. If the extracted features are carefully chosen it is expected that the feature set will extract the most relevant information from the input feature vectors. Feature extraction simplifies the amount of resources required to describe a large set of data accurately. When performing analysis of complex data major problems arise from the number of variables involved. Analysis with a large number of variables in the input feature vector requires a large amount of memory and computation power. Another major problem arises with large input feature vector is that the classification algorithm overfits the training sample and generalizes poorly for new samples those are not subset of the training samples [5]. Feature extraction is generalized terms for procedures of constructing combinations of the variables to get solve these problems while still describing the input feature vectors

with sufficient accuracy. For feature extraction the best results can be achieved when an expert of that field constructs a set of application-dependent features. If there is no such expert knowledge available, general dimensionality reduction techniques can be applied [5]. There are several dimensionality reduction techniques are available including:

1. Linear Dimensionality Reduction
2. Non-linear Dimensionality Reduction
3. Principal components analysis
4. Kernel PCA
5. Multilinear PCA
6. Multifactor dimensionality reduction
7. Semidefinite embedding
8. Multilinear subspace learning
9. Partial least squares
10. Independent component analysis
11. Latent semantic analysis
12. Isomap

3.3 Linear dimensionality reduction

Among all available techniques mentioned above, linear dimensionality reduction techniques are the preferred ones because of their linear time computation complexity and their efficiency. Linear dimensionality reduction has been studied for a long time in the field of pattern recognition.

The basic idea of LDR is to represent an object of dimension n as a lower-dimensional vector of dimension d (where $d << n$), achieving this by performing a linear transformation. We consider two classes, ω_1 and ω_2, represented by two normally distributed random vectors $\mathbf{x}_1 \sim N(\mathbf{m}_1, \mathbf{S}_1)$ and $\mathbf{x}_2 \sim N(\mathbf{m}_2, \mathbf{S}_2)$, respectively, with p_1 and p_2 are the *a priori* probabilities of these two classes. After the LDR is applied, two new random vectors $\mathbf{y}_1 = \mathbf{A}\mathbf{x}_1$ and $\mathbf{y}_2 = \mathbf{A}\mathbf{x}_2$ are generated, where $\mathbf{y}_1 \sim N(\mathbf{A}\mathbf{m}_1; \mathbf{A}\mathbf{S}_1\mathbf{A}^t)$ and $\mathbf{y}_2 \sim N(\mathbf{A}\mathbf{m}_2; \mathbf{A}\mathbf{S}_2\mathbf{A}^t)$ with $\mathbf{m}_i$ and $\mathbf{S}_i$ being the mean vectors and covariance matrices in the original space, respectively. The aim of LDR is to find a linear transformation matrix $\mathbf{A}$ in such a way that the new classes ($\mathbf{y}_i = \mathbf{A}\mathbf{x}_i$) are as separable as possible. Where $\mathbf{S}_W = p_1\mathbf{S}_1 + p_2\mathbf{S}_2$ and $\mathbf{S}_E = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t$ be the within-class and between-class scatter matrices respectively.

3.3.1 Linear discriminant analysis

There are various schemes available that yield linear dimensionality reduction. Linear discriminant analysis is one of them. Linear discriminant analysis includes methods to find a linear combination of features which characterizes or separates two or more classes of objects. Linear discriminant analysis is closely related to regression analysis which also attempts to express one dependent variable as linear combinations of other features. Linear discriminant analysis is also related to principal component analysis in that both look for

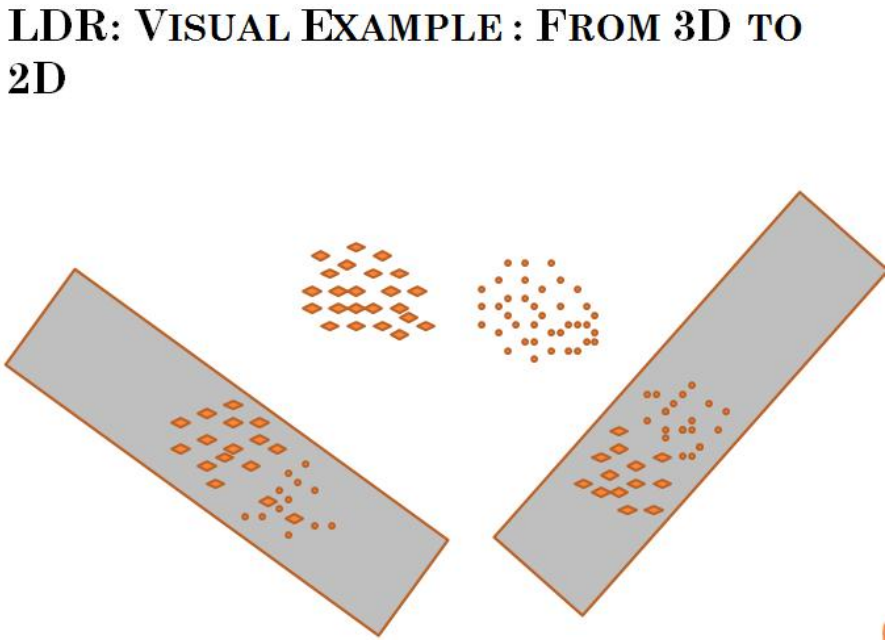

Figure 3.1: Linear Dimensionality Reduction.

linear combinations of variable which represents the original data as accurately as possible. But principal component analysis does not take into account different classes. Linear discriminant analysis is a supervised method for dimensionality reduction, if we say x and y are samples from two classes, in linear discriminant analysis we want to find the direction that is defined by vector A, such that when data are projected onto A the examples from two classes are as well separated as possible. In this thesis to reduce feature dimensions we have used three different linear discriminant analysis criteria including:

1. Fishers Discriminant Analysis
2. Heteroscedastic Discriminant Analysis
3. Chernoff Discriminant Analysis

3.3.2 Fisher's discriminant analysis

The Fisher's discriminant analysis was proposed by Ronald Fisher [5]. We assume that we have two classes ω_1 and ω_2 where $\mathbf{S}_W = p_1\mathbf{S}_1 + p_2\mathbf{S}_2$ and $\mathbf{S}_E = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t$ be the within-class and between-class scatter matrices respectively. The well-known Fisher's discriminant analysis criterion consists of maximizing the Mahalanobis distance between the transformed distributions by finding $\mathbf{A}$ that maximizes the following function [5]

$$J_{FDA}(\mathbf{A}) = tr\left\{(\mathbf{A}\mathbf{S}_W\mathbf{A}^t)^{-1}(\mathbf{A}\mathbf{S}_E\mathbf{A}^t)\right\} . \tag{3.1}$$

The matrix $\mathbf{A}$ that maximizes (3.1) is obtained by finding the eigenvalue decomposition of the matrix:

$$\mathbf{S}_{FDA} = \mathbf{S}_W^{-1}\mathbf{S}_E , \tag{3.2}$$

and taking the d eigenvectors whose eigenvalues are the largest ones. Since $\mathbf{S}_E$ is of rank one, $\mathbf{S}_W^{-1}\mathbf{S}_E$ is also of rank one. Thus, the eigenvalue decomposition of $\mathbf{S}_W^{-1}\mathbf{S}_E$ leads to only one non-zero eigenvalue, and hence FDA can only reduce to dimension $d = 1$.

3.3.3 Heteroscedastic Discriminant Analysis

HDA has been recently proposed as a new LDR technique for normally distributed classes [18], which takes the Chernoff distance in the original space into consideration to minimize the error rate in the transformed space. It can be seen as a generalization of FDA to consider heteroscedastic classes, and the aim is to obtain the matrix $\mathbf{A}$ that maximizes the function:

$$J_{HDA}(\mathbf{A}) = tr\left\{(\mathbf{A}\mathbf{S}_W\mathbf{A}^t)^{-1}\left[\mathbf{A}\mathbf{S}_E\mathbf{A}^t - \mathbf{A}\mathbf{S}_W^{\frac{1}{2}}\frac{p_1\log(\mathbf{S}_W^{-\frac{1}{2}}\mathbf{S}_1\mathbf{S}_W^{-\frac{1}{2}})+p_2\log(\mathbf{S}_W^{-\frac{1}{2}}\mathbf{S}_2\mathbf{S}_W^{-\frac{1}{2}})}{p_1p_2}\mathbf{S}_W^{\frac{1}{2}}\mathbf{A}^t\right]\right\} \tag{3.3}$$

where the logarithm of a matrix $\mathbf{M}$, $\log(\mathbf{M})$, is defined as:

$$\log(\mathbf{M}) \triangleq \Phi\log(\Lambda)\Phi^{-1}. \tag{3.4}$$

with Φ and Λ representing the eigenvectors and eigenvalues of $\mathbf{M}$, respectively.

The solution to this criterion is given by computing the eigenvalue decomposition of:

$$\mathbf{S}_{HDA} = \mathbf{S}_W^{-1}\left[\mathbf{S}_E - \mathbf{S}_W^{\frac{1}{2}}\frac{p_1\log(\mathbf{S}_W^{-\frac{1}{2}}\mathbf{S}_1\mathbf{S}_W^{-\frac{1}{2}})+p_2\log(\mathbf{S}_W^{-\frac{1}{2}}\mathbf{S}_2\mathbf{S}_W^{-\frac{1}{2}})}{p_1p_2}\mathbf{S}_W^{\frac{1}{2}}\right] \tag{3.5}$$

and choosing the d eigenvectors whose corresponding eigenvalues are the largest ones.

3.3.4 Chernoff discriminant analysis

Cheroff Discriminant Analysis criterion is a type of LDR method that has been recently proposed. The aim of the criteria is to maximize the separability of the distributions in the transformed space measured by distance between two classes. CDA assumes that the classes are normally distributed in the original and also in the transformed spaces by maximizing the following criteria [26]:

$$\begin{aligned} J_{CDA}(\mathbf{A}) = tr\{&p_1p_2\mathbf{A}\mathbf{S}_E\mathbf{A}^t(\mathbf{A}\mathbf{S}_W\mathbf{A}^t)^{-1} \\ &+\log(\mathbf{A}\mathbf{S}_W\mathbf{A}^t) - p_1\log(\mathbf{A}\mathbf{S}_1\mathbf{A}^t) - p_2\log(\mathbf{A}\mathbf{S}_2\mathbf{A}^t)\} \end{aligned} \tag{3.6}$$

where $\mathbf{S}_W = p_1\mathbf{S}_1 + p_2\mathbf{S}_2$, $\mathbf{S}_E = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t$.

In [26] it has been shown that for any normally distributed random vectors, $\mathbf{x_1}$ and $\mathbf{x_2}$, there always exists an orthogonal matrix $\mathbf{Q}$, where $\mathbf{QQ}^t = \mathbf{I}$, such that $J_{CDA}(\mathbf{A}) = J_{CDA}(\mathbf{Q})$ for any $\mathbf{A}$ or rank d. Thus we can assume that $\mathbf{A}$ is an orthogonal matrix. A gradient-based algorithm was proposed in [26] that maximizes the function (4.9) in an iterative way. The algorithm starts with an arbitrary orthogonal matrix $\mathbf{A}^{(1)}$. At the step $k+1$ the orthogonal matrix is computed as follows:

$$\mathbf{A}^{(k+1)} = \mathbf{A}^{(k)} + \alpha_k \nabla J_{CDA}(\mathbf{A}^{(k)}) \tag{3.7}$$

where the gradient for J_{CDA} is:

$$\begin{aligned}
&\tfrac{\partial J_{CDA}}{\partial \mathbf{A}} = \nabla J_{CDA}(\mathbf{A}) = 2p_1p_2\left[\mathbf{S}_E\mathbf{A}^t(\mathbf{A}\mathbf{S}_W\mathbf{A}^t)^{-1}\right.\\
&\left.-\mathbf{S}_W\mathbf{A}^t(\mathbf{A}\mathbf{S}_W\mathbf{A}^t)^{-1}(\mathbf{A}\mathbf{S}_E\mathbf{A}^t)(\mathbf{A}\mathbf{S}_W\mathbf{A}^t)^{-1}\right]^t\\
&+2\left[\mathbf{S}_W\mathbf{A}^t(\mathbf{A}\mathbf{S}_W\mathbf{A}^t)^{-1} - p_1\mathbf{S}_1\mathbf{A}^t(\mathbf{A}\mathbf{S}_1\mathbf{A}^t)^{-1}\right.\\
&\left.-p_2\mathbf{S}_2\mathbf{A}^t(\mathbf{A}\mathbf{S}_2\mathbf{A}^t)^{-1}\right]^t
\end{aligned}$$

For the above mentioned gradient algorithm let the the learning rate is α_k. To make sure that the gradient algorithm converges, the learning rate needs to be maximized. In the Rueda et. al. study [26], to maximize this function the secant method is proposed:

$$\phi_k(\alpha) = J_{CDA}(\mathbf{A}^{(k)} + \alpha\nabla J_{CDA}(\mathbf{A}^{(k)})) \tag{3.8}$$

The initial value of learning rate starts with $\alpha^{(0)}$ and $\alpha^{(1)}$. The value of the learning rate $\alpha^{(j+1)}$ at time $j+1$ is as follows:

$$\alpha^{(j+1)} = \alpha^{(j)} + \frac{\alpha^{(j)} - \alpha^{(j-1)}}{\frac{d\phi_k}{d\alpha}(\alpha^{(j)}) - \frac{d\phi_k}{d\alpha}(\alpha^{(j-1)})} \frac{d\phi_k}{d\alpha}(\alpha^{(j)}) \tag{3.9}$$

where

$$\frac{d\phi_k}{d\alpha}(\alpha) = [\nabla J_{CDA}(\mathbf{A}^{(k)} + \alpha \nabla J_{CDA}(\mathbf{A}^{(k)}))] \cdot \nabla J_{CDA}(\mathbf{A}^{(k)}) \tag{3.10}$$

The dot product between two matrices is represented by the operator "$\cdot$". It is computed for any two matrices $\mathbf{C}$ and $\mathbf{D}$, as $\mathbf{C} \cdot \mathbf{D} = tr\{\mathbf{C}\,\mathbf{D}\}$. By replacing $\mathbf{A}$ for $(\mathbf{A} + \alpha \nabla J_{CDA}(\mathbf{A}))$ in the equation (3.8) the value of $\nabla J_{CDA}(\mathbf{A}^{(k)} + \alpha \nabla J_{CDA}(\mathbf{A}^{(k)}))$ is computed.

With the definition of $\frac{d\phi_k}{d\alpha}(\alpha)$, equation (3.9) can be solved, and the gradient algorithm continues with the next iteration. The full complete algorithm is found in [26]. One of the key aspects of the algorithm is the initialization of the matrix $\mathbf{A}$. In this thesis for our study we have performed ten different random initializations and then have chosen the solution for $\mathbf{A}$ which yields the maximum Chernoff distance.

3.4 Classifier

In pattern recognition classification is the procedure of identifying unknown new samples on the basis of training set of known samples. The classification problem is known as

supervised learning in pattern recognition. In supervised learning, the algorithm analyzes the training data and produces an inferred function. This inferred function; if its output is discrete it is called classifier. This classifier should predict the correct output value for any valid input object.

To do the classification classifiers are fed by training data. The training data set needs to be representative of the real-world use of the function. The accuracy of the classifier depends on how the training data is represented. If the training dataset is too large, it will face the problem of the curse of dimensionality [5] and for that reason we have to pre-process the training data with feature extraction methods as explained above. On the basis of the representation of the input data the algorithm will generate the learned function. After that, the accuracy of the generated learned function is evaluated on a test set that is separate from the training set. There are several types of classifiers available [5]:

1. Linear classifiers
2. Quadratic classifiers
3. Support Vector Machines
4. Kernel estimation
5. Decision trees
6. Neural Networks
7. Hidden Markov Models

Among these for our study we have used support vector machine, linear classifier and quadratic classifier.

3.4.1 Support vector machine

SVM is a supervised learning algorithm in machine learning that constructs a hyperplane or a set of hyperplanes in a high dimensional space which can be used as a separator to separate different classes. Let x_i where i =1,2,3.....n, be the feature vectors of the training set X. These belong to either of the two classes ω_1 and ω_2, which are assumed to be linearly separable. The goal of the SVM is to design a hyperplane that classifies all the training vectors

$$g(x) = \omega^T x + \omega_0 \tag{3.11}$$

This kind of hyperplane is not unique; the simple form perceptron algorithm may converge to any one of the possible solutions. But the SVM chooses the best classifier that has least risk of causing an error when operating with unknown data. This is known as the generalization of the performance of the classifier [5]. It chooses the hyperplane that leaves the maximum margin from both classes. The distance of a point from a hyperplane is given by

$$z = \frac{|g(x)|}{||\omega||} \tag{3.12}$$

If for each x_i we denote the corresponding class indicator by y_i (+1 for ω_1, -1 for ω_2). The SVM algorithm find the best hyperplane by computing the parameters ω, ω_0 of the hyperplane so that to minimize

$$J(\omega) = \frac{1}{2}||\omega||^2 \tag{3.13}$$

subject to

$$y_i(\omega^T x_i + \omega_0) \geq 1, i = 1, 2, ...N \tag{3.14}$$

SVM can also be tuned to do a non-linear classification by applying the kernel trick to maximum-margin hyperplanes. The effectiveness of SVM depends on the selection of the kernel, the selection parameters and the soft margin [5].

3.4.2 Quadratic classifier

In Pattern Recognition Quadratic classifier is an algorithm used to separate two classes of objects by a quadratic boundary. The surface of separating two classes for a quadratic classifier would be a conic section (a line, a circle, an ellipse, a parabola, a hyperbola). Bayesian classifier for normally distributed classes is an example of quadratic classifier [5]. One of the most popular probability density functions in practice is the Gaussian or normal density function, because of its computational tractability and the fact that it models

adequately a large number of cases. Now we are assuming that the likelihood functions of the ω_i with respect to x in the n-dimensional feature space follows the general multivariate normal density [5].

$$p(x|\omega_i) = \frac{1}{(2\pi)^{n/2}|\Sigma_i n^{\frac{1}{2}}} exp(-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-n}(x-\mu_i)) \tag{3.15}$$

$i = 1,...,M$, where μ_i = $E[x]$ is the mean value of the ω_i class and Σ_i is the $n \times n$ covariance matrix. $|\Sigma_i|$ denotes the determinant of Σ_i and $E[.]$ the mean value of a random variable. Sometimes, the symbol $\eta(\mu,\Sigma)$ is used to denote a Gaussian probability density function with mean value μ and covariance Σ. Now the discriminant functions of the Bayesian classifier can be achieved in a nonlinear quadratic form. For example the case of $n = 2$ the discriminant functions becomes [5]

$$g_i(x) = -\frac{1}{2\sigma_i^2}(x_1^2 + x_2^2) + \frac{1}{\sigma_i^2}(\mu_i 1 x_1 + \mu_i 2 x_2) - \frac{1}{2\sigma_i^2}(\mu_i^2 1 + \mu_i^2 2) + lnP(\omega_i) + c_i \tag{3.16}$$

and the decision curves $g_i(x) - g_j(x) = 0$ are quadrics (ellipsoids, parabolas, hyperbolas, pair of lines). In this case, the Bayesian classifier is a quadratic classifier, in the sense that the partition of the feature space is performed via quadric decision surface [5].

3.4.3 Linear classifier

Linear classifier is nothing but a specialized form of the above mentioned quadratic classifier where the decision boundary is achieved by a straight line. Perceptron, naive bayes classifier, logistic regression, Fisher's linear discriminant analysis are different examples of linear classifiers. For the above mentioned Quadratic Bayesian classifier when $g_i(x)$ is a linear function of x, then that Bayesian classifier is a linear classifier.

3.5 m-fold cross validation

In statistical pattern recognition cross-validation is a method for assessing the accuracy of the predictive model will perform in practice. Every fold of cross validation involves partitioning the data into different subsets, use one part of the subset as the training set for the classifier and the other part of the subset as the testing set for the classifier. To reduce the variance we have performed ten rounds of cross using different partitions. Thus, the dataset is randomly partitioned in ten subsets. Among these ten subsets one subset is used for testing and the other remaining nine subsets are used as training data. This procedure is repeated ten times with each of these ten subsets is used exactly once as testing data. Then, we have averaged these ten results to produce a single evaluation. The advantage of using ten fold cross validation is that all subsets are used for both testing and training, and each subset is used for testing exactly once.

3.6 Prediction evaluation

In pattern recognition the confusion matrix is a matrix in supervised learning. In this matrix, each column of the matrix represents the instances in a predicted class, while each row

represents the instances in an actual class. We have created the confusion matrix for each prediction, and from that confusion matrix we have created different measurements such as Accuracy, Specificity and Sensitivity to evaluate our prediction. We have computed the accuracy of the prediction by this formula [5]

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{3.17}$$

We have computed Specificity by this formula [5]

$$SpecificityorTrueNegativeRate = \frac{TN}{FP+TN} \tag{3.18}$$

For calculating of Sensitivity from the Confusion Matrix we have used this formula [5]

$$SensitivityorTruePositiveRate = \frac{TP}{TP+FN} \tag{3.19}$$

Where TP is the number of True Positives, TN is the number of True Negatives, FP is the number of False Positives and FN is the number of False Negatives. For the case of biological-crystal packing classification, TP is the number of correctly classified biological samples and TN is the correctly identified crystal packing samples. For the case of obligate-

nonobligate classification TP is the correctly identified obligate complex and TN is the correctly identified nonobligate complex.

3.7 Receiver operating characteristic curve

In statistical pattern recognition a receiver operating characteristic curve is a graphical plot of True Positive Rate (TPR) vs False Positive Rate (FPR) for a binary classifier system. True Positive Rate is known as sensitivity and False Positive Rate is known as $(1-Specificity)$. True Positive Rate of a classifier is calculated from the confusion matrix by this formula [5]

$$TruePositiveRateorSensitivity = \frac{TP}{TP+FN} \tag{3.20}$$

and False Positive Rate of a classifier is calculated from confusion matrix by this formula [5]

$$FalsePositiveRateor(1-Specificity) = \frac{FP}{FP+TN} \tag{3.21}$$

In the receiver operating characteristic curve the True Positive Rate is plotted along the Y-axis and the False Positive Rate is plotted along X-axis. The whole space in the graph is called receiver operating characteristic space. The receiver operating characteristic space is divided by a diagonal. A curve above diagonal represents a good classification result, any point below diagonal represents a poor classification results. True Positive Rate determines a classifier's performance on classifying positive instances correctly among all positive samples available. False Positive Rate determines a classifier's performance on

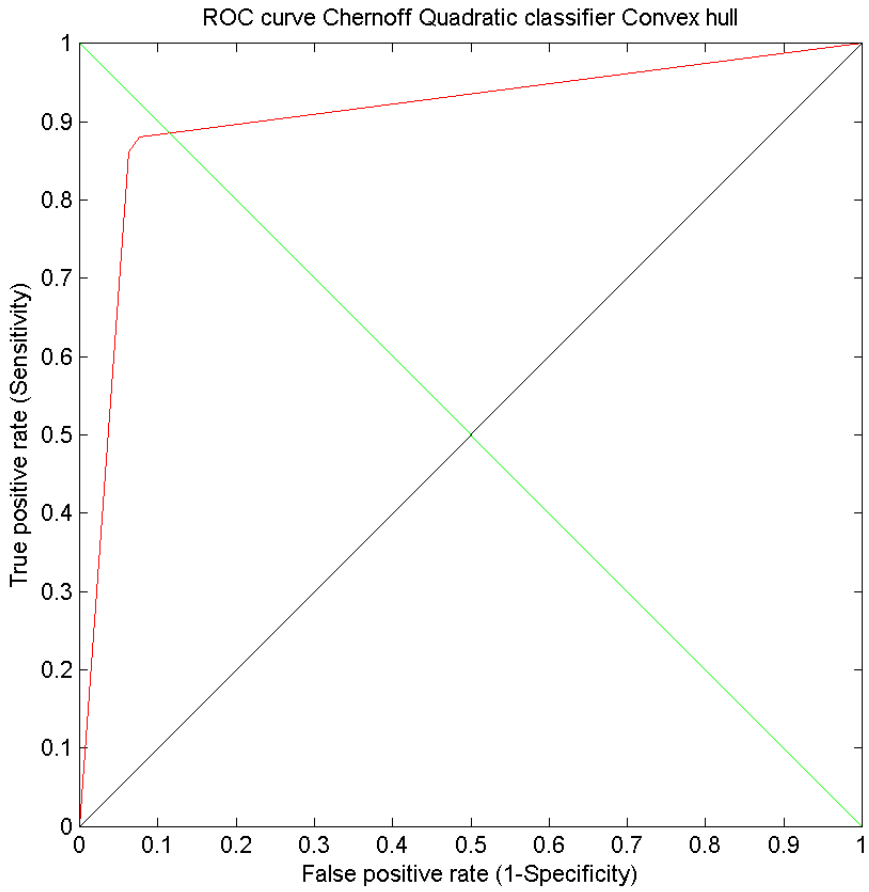

Figure 3.2: Receiver Operating Characteristic Curve for Biological-NonBiological classification with CDA Quadratic Classifier for 44 features

classifying incorrect positive results among all negative samples. The perfect prediction would yield a point in the upper left corner or coordinate $(0, 1)$ of the Receiver Operating Characteristic space. This kind of prediction represents 100 percent sensitivity that means no False Negatives and also 100 percent specificity that means no False positives.

3.8 Matthews' correlation coefficient

In pattern recognition Matthews' correlation coefficient is used as a measure of the quality of a binary classification. It considers true and false positives and negatives and is generally considered as a balanced measure which can be used even the two classes are of different sizes. It is a correlation coefficient between the observed and predicted binary classifications. The range of values of the Matthews correlation coefficient is between -1 to $+1$, where a value of $+1$ represents perfect prediction, 0 is considered as an average random prediction and -1 as an inverse prediction. The Matthews correlation coefficient can be calculated directly from the confusion matrix by using this formula [5]

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{3.22}$$

Where TP is the number of True Positives, TN is the number of True Negatives, FP is the number of False Positives and FN is the number of False Negatives. While there is no perfect way of describing the confusion matrix by a single value, the Matthews Correlation Coefficient is considered as being one of the best such measures. Another kind of measurement such as proportion of correct prediction that we know as accuracy are not very useful when the two classes are of very different sizes.

3.9 Multi class classification

In Pattern Recognition multiclass classification is a statistical classification of classifying more than two different classes or events. In this study we have used multi-staged LDR to

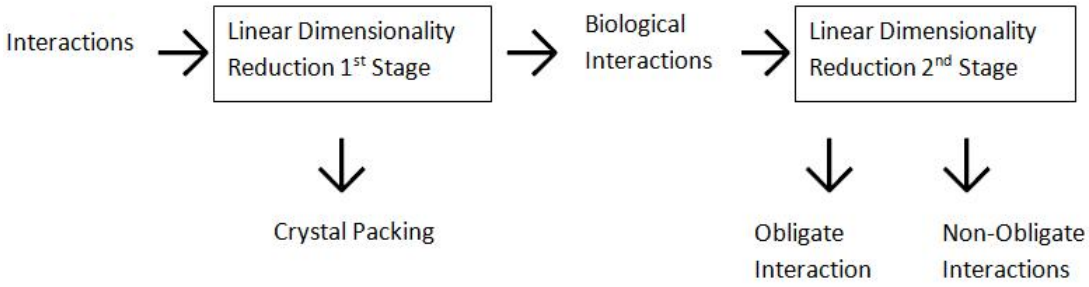

Figure 3.3: Multi class classification.

predict three different types of classes. In this study we have a 3 class problem; the three classes are crystal packing interactions, obligate interactions, non-obligate interactions. To successfully predict these 3 classes we have used multi-stage linear dimensionality reduction. In the first stage of multi-stage we have discriminated between two classes' biological interactions with crystal packing interactions. In the second stage of multi-stage we have discriminated between obligate interactions with non-obligate interactions.

Chapter 4

Methodology

4.1 Procedure for feature generation

To predict different types of protein complexes that participate in different types of interactions, we have first generated their physio-chemical features by different methods. We have generated different features from the protein-protein interaction's interface properties. Among these generated features the first six features were same as the Zhu et. al. [35] study. We have also proposed 40 new derived features from the interface property. They are number based amino acid composition and area based amino acid composition.

4.2 Calculation of interface area and interface area ratio

After automatically downloading protein three-dimensional structure files from Protein Data Bank database, we have executed NACCESS [7], the input of NACCESS is the PDB file and the corresponding PDB ID and the interacting chain names. We have automated the procedure from downloading to NACCESS computation. NACCESS [7] calculates

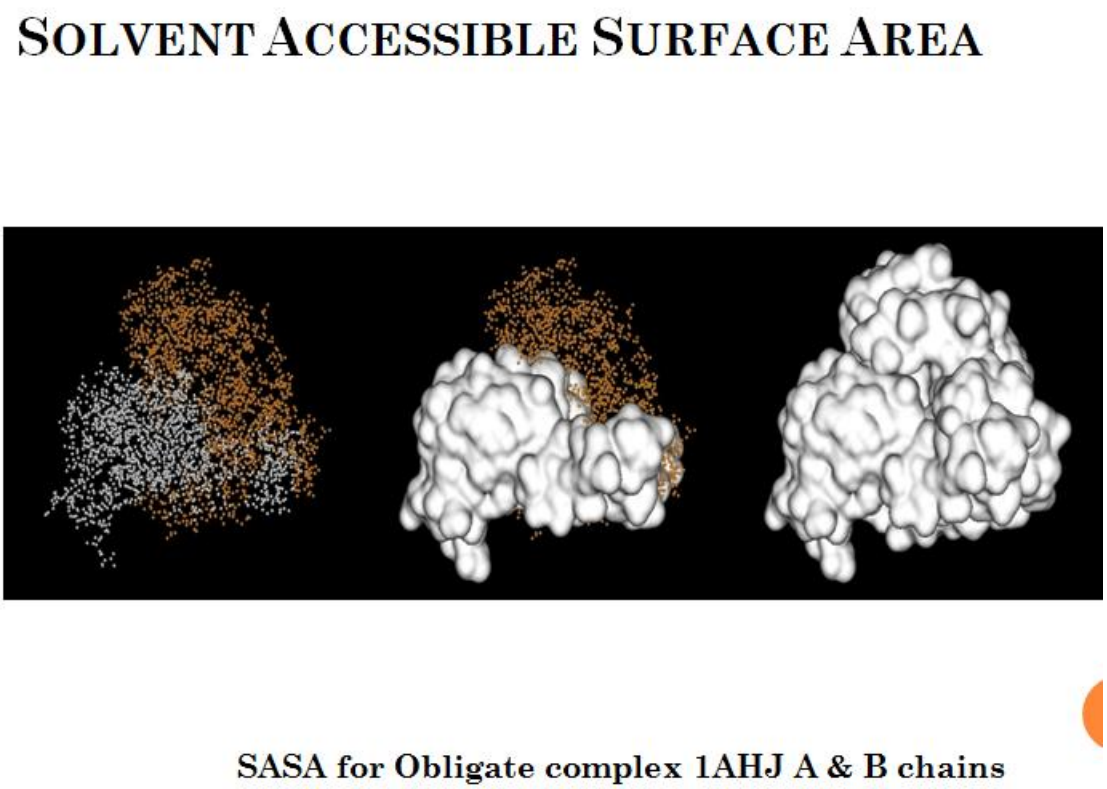

Figure 4.1: Solvent Accessible Surface Area (SASA) diagram of 1AHJ A B. Diagram prepared through GRASP, a molecular visualization package.

the Solvent Accessible Surface Area (SASA) of each residues within the protein complex, NACCESS outputs the calculated result in RSA files. SASA values are calculated to obtain the interface area of protein-protein interactions. A residue is considered as being part of the interface if its SASA value decreases by more than 1 $\AA^2$ upon the formation of the complex [30]:

From the SASA value that is outputted by NACCESS we have calculated the first feature of NOXClass [35] that is Interface Area by this formula [35]:

$$InterfaceArea = \frac{1}{2}(SASA_a + SASA_b - SASA_{ab}) \tag{4.1}$$

and the second feature Interface Area Ratio by this formula [35]:

$$InterfaceAreaRatio = \frac{InterfaceArea}{min(SASA_a, SASA_b)} \tag{4.2}$$

4.3 Proposed 40 new features

We are proposing 40 new derived features from the interface property of different protein-protein interactions. There are several studies have been done previously to differentiate obligate and non-obligate protein-protein interactions on the basis of protein-protein interaction interface properties. Among them NOXClass [35] had taken into account the amino acid composition of the interface. But they had not taken into account different *types* of amino acid composition of the protein-protein interaction interface such as number based amino acid composition and area based amino acid composition.

Table 4.1: 20 standard amino acids

Alanine	ALA
Arginine	ARG
Asparagine	ASN
Aspartic acid	ASP
Cysteine	CYS
Glutamic acid	GLU
Glutamine	GLN
Glycine	GLY
Histidine	HIS
Isoleucine	ILE
Leucine	LEU
Lysine	LYS
Methionine	MET
Phenylalanine	PHE
Proline	PRO
Serine	SER
Threonine	THR
Tryptophan	TRP
Tyrosine	TYR
Valine	VAL

4.3.1 Number based amino acid composition of the interface

The number based amino acid composition of the interface is defined as the frequency of each type of 20 standard amino acids in the protein-protein interface. Amino acids have critical function to life such as metabolism, and they also serve as the building blocks of the proteins. Amino acids can be linked together in varying sequences to form a vast variety of proteins. Twenty amino acids are naturally incorporated into polypeptides and are called standard amino acids. These 20 standard amino acids and their three letter acronyms that are mentioned in their corresponding PDB file to show the structure of the protein complex are shown in Table:4.1.

To calculate number based amino acid composition or the frequency of each type of the above mentioned 20 standard amino acids in the protein-protein interaction interface, we first downloaded the PDB files of the protein complex from the Protein Data Bank (http://www.pdb.org/pdb/home/home.do) website. The data stored in the Protein Data Bank database is typically obtained by X-ray Crystallography and NMR Spectroscopy and sub-

mitted by biochemists. Currently, Protein Data Bank has structural data of more than 68000 proteins (ref: http://www.rcsb.org/pdb/statistics/holdings.do). There are three kinds of format files are available from Protein Data Bank .PDB format .mmCIF(macromolecular Crystallographic Information file) format and .PDBML(XML version) format. We have downloaded .PDB file format, this file format has 80 characters per line. Each protein structure published in PDB receives a four character alphanumeric identifier. We have applied our model of classification on two different collections (namely Zhu dataset [35] and Mintseris dataset [20]) of this type of alphanumeric characters.

After downloading protein structure files from PDB, we have executed NACCESS [7], the input of NACCESS is the PDB file and the corresponding PDB ID and the interacting chain names. NACCESS [7] calculates the Solvent Accessible Surface Area values of each residues within the protein complex, NACCESS outputs the calculated result in RSA files. From these RSA files we have calculated the frequency of each 20 standard amino acid in the protein-protein interaction interface, and that results in 20 features of that protein-protein interaction. These features are small integers for our experimented datasets.

4.3.2 Area based amino acid composition of the interface

By weighting each residue with it's Δ SASA, we have calculated the area based amino acid compositions. It is also calculated from the RSA file. For 20 amino acids we have got 20 area based amino acid composition values. The calculation of 20 standard amino acids can be obtained by this formula:

$$v_{a,i=1,\ldots,20} = \frac{1}{2\,Interface\,Area}\Sigma_{r,type(r)=i}\Delta SASA(r) \tag{4.3}$$

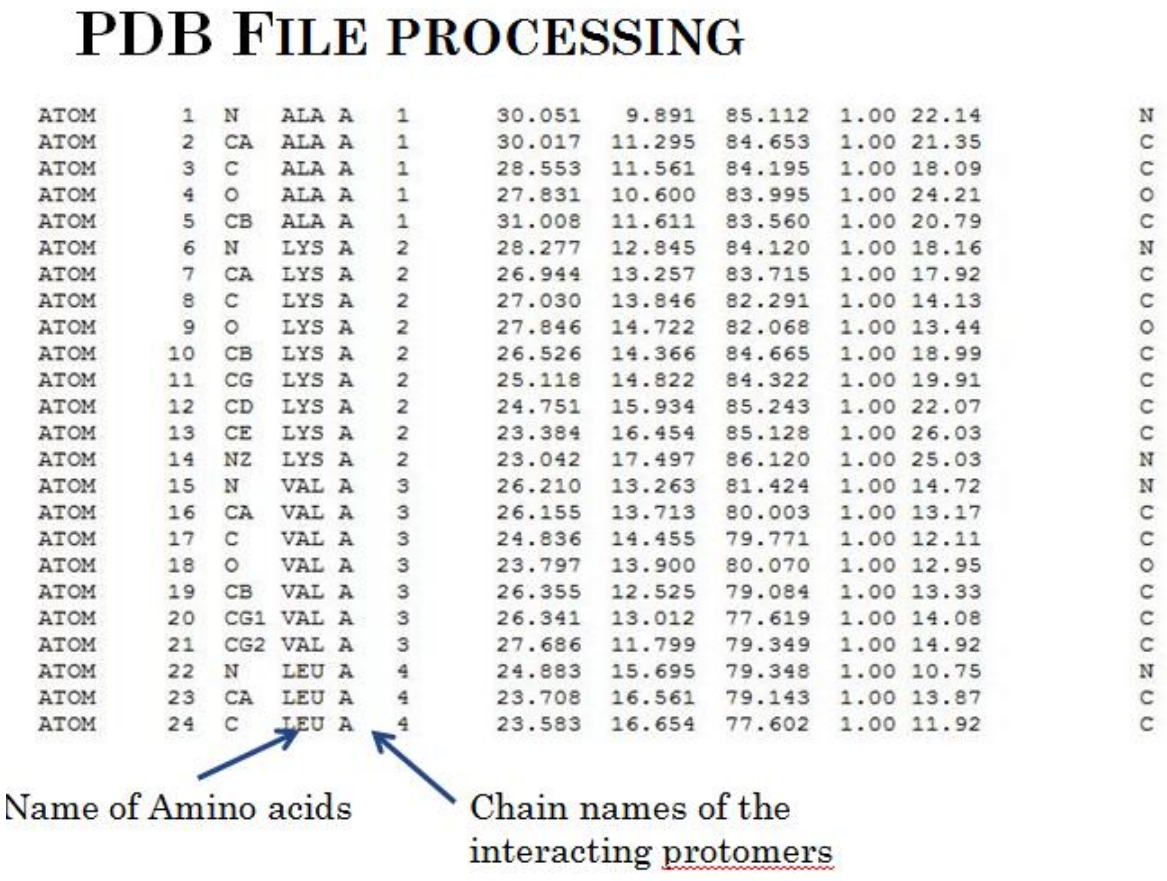

Figure 4.2: PDB file for 1AHJ A B downloaded from PDB.

RSA FILE PROCESSING

```
REM  Relative accessibilites read from external file "standard.data"
REM  File of summed (Sum) and % (per.) accessibilities for
REM RES _ NUM      All-atoms   Total-Side   Main-Chain     Non-polar     All polar
REM                  ABS   REL    ABS   REL    ABS   REL    ABS   REL    ABS   REL
RES ALA A   1     53.32  49.4  24.77  35.7  28.56  74.1  28.37  39.7  24.96  68.2
RES LYS A   2     29.40  14.6  29.40  18.0   0.00   0.0  21.27  18.2   8.13   9.6
RES VAL A   3      0.00   0.0   0.00   0.0   0.00   0.0   0.00   0.0   0.00   0.0
RES LEU A   4      0.00   0.0   0.00   0.0   0.00   0.0   0.00   0.0   0.00   0.0
RES CYS A   5      0.00   0.0   0.00   0.0   0.00   0.0   0.00   0.0   0.00   0.0
RES VAL A   6      0.00   0.0   0.00   0.0   0.00   0.0   0.00   0.0   0.00   0.0
RES LEU A   7      0.00   0.0   0.00   0.0   0.00   0.0   0.00   0.0   0.00   0.0
RES TYR A   8     39.78  18.7  39.78  22.4   0.00   0.0  37.99  27.8   1.79   2.3
RES ASP A   9     89.79  64.0  78.45  76.4  11.34  30.1  29.97  60.9  59.82  65.6
RES ASP A  10     37.69  26.8  25.40  24.7  12.29  32.6  12.24  24.9  25.45  27.9
RES PRO A  11     47.74  35.1  43.11  36.0   4.63  28.6  43.11  35.6   4.63  30.5
RES VAL A  12    161.16 106.4 127.39 111.5  33.77  90.9 128.32 111.1  32.84  91.3
RES ASP A  13    134.64  95.9 102.54  99.9  32.10  85.2  54.73 111.1  79.91  87.7
RES GLY A  14     35.71  44.6  29.29  90.6   6.42  13.4  29.82  79.4   5.89  13.8
RES TYR A  15     66.29  31.2  49.95  28.2  16.35  46.2  48.34  35.4  17.95  23.5
RES PRO A  16     37.89  27.8  32.32  27.0   5.56  34.3  32.32  26.7   5.56  36.6
RES LYS A  17    192.98  96.1 164.58 100.8  28.40  75.7 115.81  99.3  77.17  91.6
RES THR A  18    113.30  81.4 100.54  98.9  12.76  34.0  70.09  92.6  43.21  68.0
RES TYR A  19     72.63  34.1  43.99  24.8  28.65  81.0  47.31  34.7  25.32  33.2
RES ALA A  20    114.10 105.7  75.31 108.5  38.79 100.7  76.08 106.6  38.02 103.9
RES ARG A  21    147.46  61.8 135.55  67.4  11.92  31.8  57.41  73.8  90.05  55.9
RES ASP A  22    142.61 101.6 116.95 113.9  25.66  68.1  44.65  90.7  97.96 107.5
RES ASP A  23     89.95  64.1  89.41  87.1   0.53   1.4  36.71  74.6  53.24  58.4
RES LEU A  24     39.04  21.9  19.16  13.6  19.88  53.0  19.60  13.8  19.44  53.5
```

Number based amino acid composition: Frequency of each type of the 20 standard amino acids

This is only Chain A, there are other with Chain B and AB

Figure 4.3: RSA file for 1AHJ A B, output of NACCESS.

After calculation of 20 number based amino acid composition values of the interface and 20 area based amino acid composition values of the interface from RSA file, we have calculated Δv distance between these two vectors where Δv = v -v' for each residue by this formula [35]:

$$(\Delta v)^2 = \frac{1}{19}\Sigma_{i=1}^{20}(v_i - v_i\prime)^2 \tag{4.4}$$

4.4 Calculation of Pearson's correlation coefficient

In statistics, Pearson's correlation coefficient between two variables X and Y is defined as the covariance of the two variables divided by the product of their standard deviations.

$$r = \frac{\Sigma_n^1(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\Sigma_n^1(X_i - \bar{X})^2}\sqrt{\Sigma_n^1(Y_i - \bar{Y})^2}} \tag{4.5}$$

In their study [34], Ofran et. al. showed that the amino acid composition of the biological interface to be significantly different from that of the rest of the protein surface. We can expect that amino acid composition of the crystal packing interface to be similar

that of the rest of the protein surface. Thus, it can be a good distinguishable feature to discriminate between biological and non-biological interaction types. For that reason we have calculated the Pearson correlation coefficient of amino acid composition of the interface and the protein surface. To calculate the coefficient we have first calculated the amino acid composition of the interface and then amino acid composition of the surface by the above mentioned procedure.

4.5 Calculation of conservation score of the interface

To calculate the Conservation score of the interface we have downloaded .grades file from the Consurf-DB. ConSurf-DB provides evolutionary conservation profiles for proteins of known structure in the PDB. Amino acid sequences similar to each sequence in the PDB were collected and multiply aligned using PSI-BLAST and MUSCLE, respectively. The evolutionary conservation of each amino acid position in the alignment was calculated using the Rate4Site algorithm, implemented in the ConSurf web-server. The algorithm takes explicitly into account the phylogenetic relations between the aligned proteins and the stochastic nature of the evolutionary process. Rate4Site assigns a conservation level for each residue using empirical Bayesian inference.

For each protein complex and for each chain, we have calculated the conservation score of each complex. After obtaining the conservation score we have weighted conservation score of each residue by its Δ SASA. After that we have calculated the area based conservation score of the interface by taking average of these weighted residue conservation score.

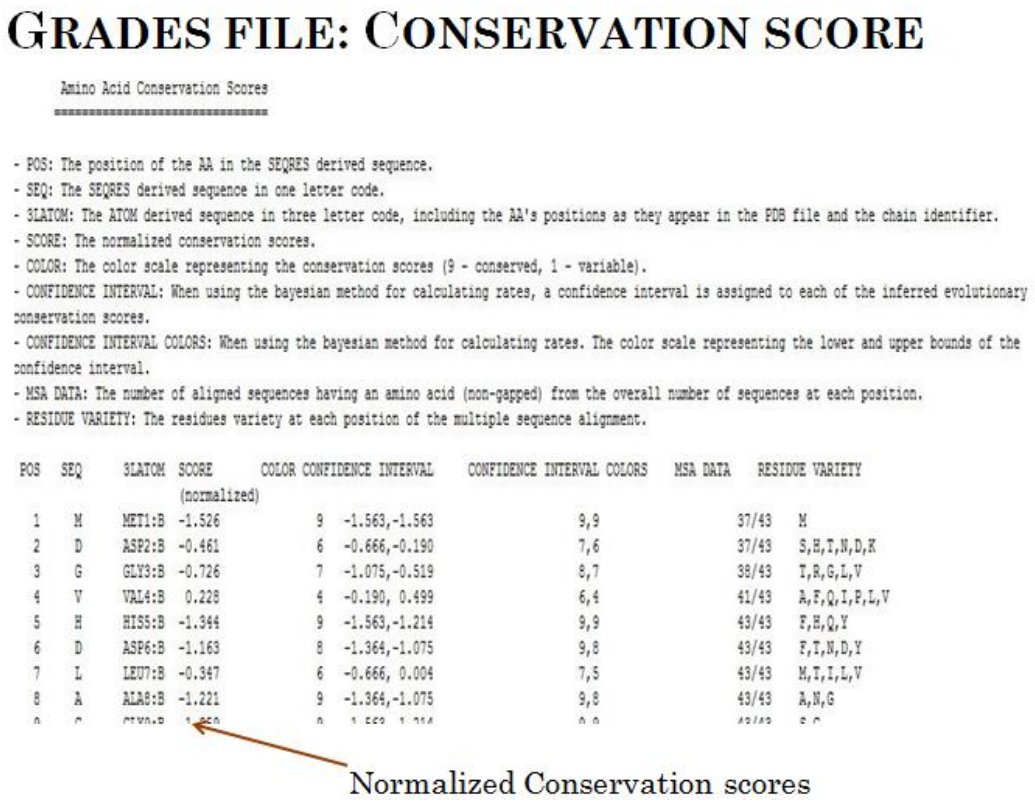

Figure 4.4: Conservation score file or .grades file of the complex 1AHJ A B downloaded from Consurf DB.

4.6 Crystal packing contacts feature generation

We have received PDB files for crystal packing interactions from Dr. Hongbo Zhu (hzhu@mpi-sb.mpg.de) of Max-Planck Institute of Germany. They were divided into "MODEL 0" and "MODEL 1" format; in these two different models they had two different interacting chains of the protein complex. To calculate SASA values of these protein complexes we had to run NACCESS on them. As the input of NACCESS is a single PDB file, we had to merge these two different models into a single file.

After merging those two different model files we ran NACCESS on them to obtain SASA values and calculated interface area and interface area ratio property as the previously mentioned procedure. We also obtain number based amino acid composition, area based amino acid composition and amino acid composition features by applying the previously mentioned procedure. We also calculated Pearson's correlation coefficient of amino acid composition of the interface and of the protein surface. The conservation score of the interface property was not calculated for crystal packing contacts as these crystal packing complexes do not have any .grades files in Consurf-DB server. Thus, finally we have computed 44 features for each crystal packing complex.

4.7 Singularity problem

As we have mentioned previously that LDR is used to reduce a dataset of large dimension to a lesser dimension dataset by making sure that the two classes are as separable as possible. This feature extraction procedure is applied prior to classification to make sure that the classifier does not suffer from the problem of curse of dimensionality. There are different LDA criteria that we have explained before to reduce the dimension of the dataset.

But when the original dimensions of feature vector of the LDR is very large the columns of the feature vectors become linearly dependent on each other resulting lower rank matrices. In linear algebra, the column rank of a matrix is found by calculating maximum number of linearly independent column vectors in that matrix. If the features for samples in our dataset or the columns of our feature vectors become large the dependency between the columns increases. That results in a lower rank matrix, and this low rank matrix creates nearly singular matrices during the calculation of the CDA criteria [26].

$$\begin{aligned} J_{CDA}(\mathbf{A}) = tr\{ & p_1 p_2 \mathbf{A}\mathbf{S}_E\mathbf{A}^t(\mathbf{A}\mathbf{S}_W\mathbf{A}^t)^{-1} \\ & + \log(\mathbf{A}\mathbf{S}_W\mathbf{A}^t) - p_1\log(\mathbf{A}\mathbf{S}_1\mathbf{A}^t) - p_2\log(\mathbf{A}\mathbf{S}_2\mathbf{A}^t)\} \end{aligned} \tag{4.6}$$

A square matrix that does not have a matrix inverse is called a singular matrix. A matrix is singular if and only if its determinant is 0. In our case the matrix is not singular, but is nearly singular. That means that the determinant is not 0 but it is very near 0 and a very small number. If we want to found the inverse of this nearly singular matrix it would generate complex numbers in the reduced feature vector. If we want to feed this reduced feature vector that contains complex numbers to our classifier, our classifier would produce unexpected results.

In our protein-protein interaction study we have encountered very large input feature vectors that contains even up to 646 features. When we ran our Linear Dimensionality Reduction to these kinds of large datasets we have encountered nearly singular matrix generation in the middle of the calculation and we have encountered complex numbers in reduced feature set and consequently erratic behavior from our classifier.

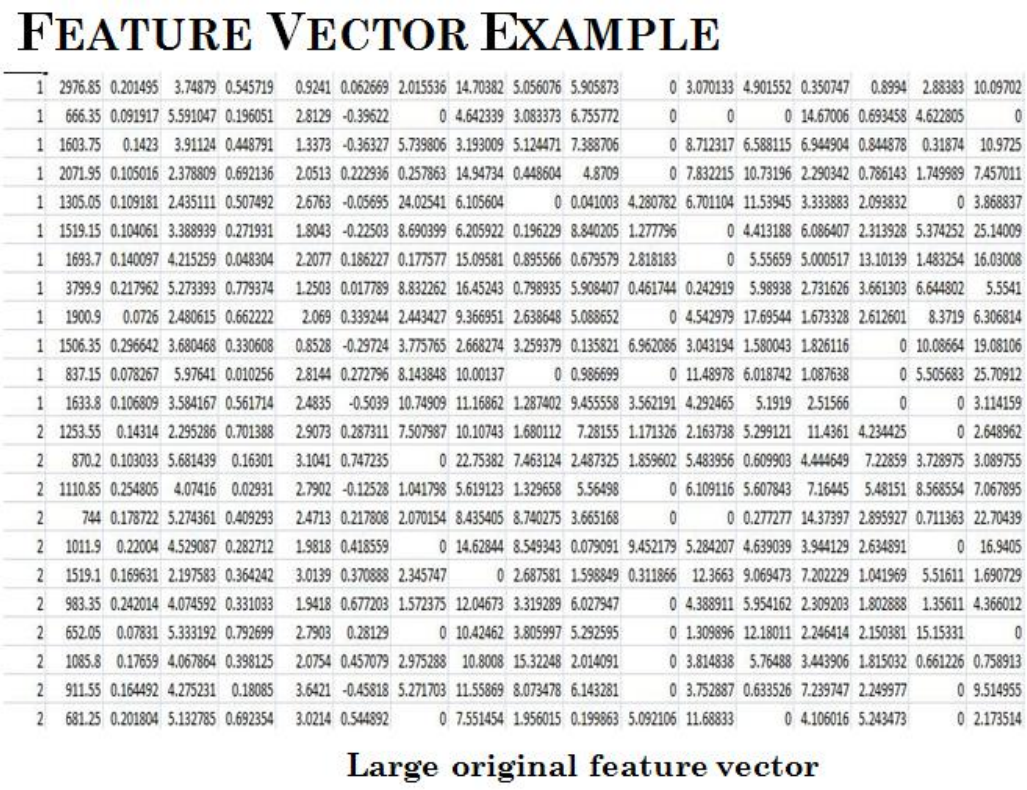

Figure 4.5: An example of large feature vector. In protein-protein interaction we have encountered up to 646 features.

4.8 Proposed solution for the Singularity problem

To solve the problem of this nearly singular matrix generation and subsequent effects of that, we applied singular value decomposition of each matrix of which we have to take the inverse of and substitute those (those values are lower than the threshold) singular values with our threshold. After replacing the values with threshold we have taken the inverse of the matrix.

4.8.1 Singular Value Decomposition

In linear algebra singular value decomposition is a factorization of the real or complex matrix. The singular value decomposition of a $m \times n$ real matrix M is a factorization of the form

$$M = U\Sigma V' \tag{4.7}$$

1. U is a mXm real or complex unitary matrix
2. Σ is a mXn diagonal matrix with non-negative real numbers on the diagonal
3. V' that is a conjugate transpose of V is a nXn real or complex unitary matrix

In the above representation U is a $m \times m$ real or complex unitary matrix, Σ is a $m \times n$ diagonal matrix with non-negative real numbers on the diagonal and V' that is a conjugate transpose of V is a $n \times n$ real or complex unitary matrix. The diagonal values Σ_i, j in the

matrix Σ are known as the singular values of the matrix M. The m columns of U and n columns of V are called the left singular vectors and the right singular vectors of the matrix M respectively.

Singular value decomposition has lots of applications in linear algebra. Among them the most popular one is that they can be used to compute the pseudo inverse of a matrix. The pseudo inverse of the matrix M is

$$M^{+} = U\Sigma^{+}V' \tag{4.8}$$

Where is Σ^{+} is the pseudo inverse of Σ, which is formed by replacing every nonzero diagonal entry by its reciprocal and transposing the resulting matrix. For our problem M is a square matrix and for that reason U and V are square too. We have used this feature of singular value decomposition to address the singularity matrix issue of linear dimensionality reduction.

4.8.2 Linear dimensionality reduction- singular value decomposition

As per the proposed solution for the nearly singular matrices problem with Linear Dimensionality Reduction we have first decomposed all the matrices of those we have to take an inverse of in this CDA criterion.

$$J_{CDA}(\mathbf{A}) = tr\{p_1 p_2 \mathbf{A}\mathbf{S}_E \mathbf{A}^t (\mathbf{A}\mathbf{S}_W \mathbf{A}^t)^{-1} + \log(\mathbf{A}\mathbf{S}_W \mathbf{A}^t) - p_1 \log(\mathbf{A}\mathbf{S}_1 \mathbf{A}^t) - p_2 \log(\mathbf{A}\mathbf{S}_2 \mathbf{A}^t)\} \tag{4.9}$$

We have decomposed all of the matrices that we have to take the inverse of. Decomposing the matrix generated the right singular vector, left singular vector and the singular values of the matrix. Here is an example of how a singular value of a matrix that we have dealt with look like.

$$\Sigma = \begin{vmatrix} .5\times10^{-2} & 0 & 0 & 0 & 0 \\ 0 & .5\times10^{-11} & 0 & 0 & 0 \\ 0 & 0 & .5\times10^{-22} & 0 & 0 \\ 0 & 0 & 0 & .5\times10^{-30} & 0 \\ 0 & 0 & 0 & 0 & .5\times10^{-45} \end{vmatrix}.$$

Now we have set the threshold of this singular value matrix by dividing the 1st element of the matrix by 10^{-10}. The purpose of setting up the threshold is that we can replace all values in the singular matrix that are less than threshold by the threshold. This is because when we are going to take the pseudo inverse of the matrix we have to take reciprocal of all the singular values of the matrix. If the singular values of that matrix are near zero or very small the reciprocal of them is going to yield near infinity values. In computer binary language that is an IEEE Not A Number (NAN) value. To avoid that problem we have replaced all the nonzero diagonals of the singular matrix that are less than our threshold by the threshold value. Thus our purpose here is that to use an acceptable singular value that computer can do reciprocals and would not generate NAN value as a result of the

reciprocals of the value. To set up the acceptable threshold, we have tried with different values and taken the first value of the singular value matrix and divide it by 10^{-10}. This has produced the most generic and efficient solution for our classification system.

Thus after replacing the nonzero diagonal values those are less than our threshold with the threshold that is the first value divided by 10^{-10} will look like as follows:

$$\Sigma = \begin{vmatrix} .5 \times 10^{-2} & 0 & 0 & 0 & 0 \\ 0 & .5 \times 10^{-11} & 0 & 0 & 0 \\ 0 & 0 & .5 \times 10^{-12} & 0 & 0 \\ 0 & 0 & 0 & .5 \times 10^{-12} & 0 \\ 0 & 0 & 0 & 0 & .5 \times 10^{-12} \end{vmatrix}.$$

In the above mentioned matrix we can see as the second singular value in the singular value matrix is not less than our threshold that is $.5^{12}$, for that reason it was not replaced. But the third, fourth, fifth singular values in the singular value matrix were less than our accepted threshold. If they were left untreated, we would have taken the pseudo-inverse of the matrix their reciprocals would have cause a generation of NAN values. Thus, we have replaced them with our threshold value that is *not* as near to zero as these values were previously, by making sure that the reciprocals of these values would not generate NAN values.

The singular value decomposition of an unitary matrix have a computational complexity of $O(n^{\omega})$ where ω's value is greater than 2 and less than 2.376. So the computational complexity of singular value decomposition is $O(n^2.376)$. As we know singular value decomposition is the most computationally expensive step of different Linear Discriminant Analysis criteria, we can say the computational complexity of different LDA criteria are $O(N^2.376)$. In the previously mentioned singularity problem solution we are only replacing

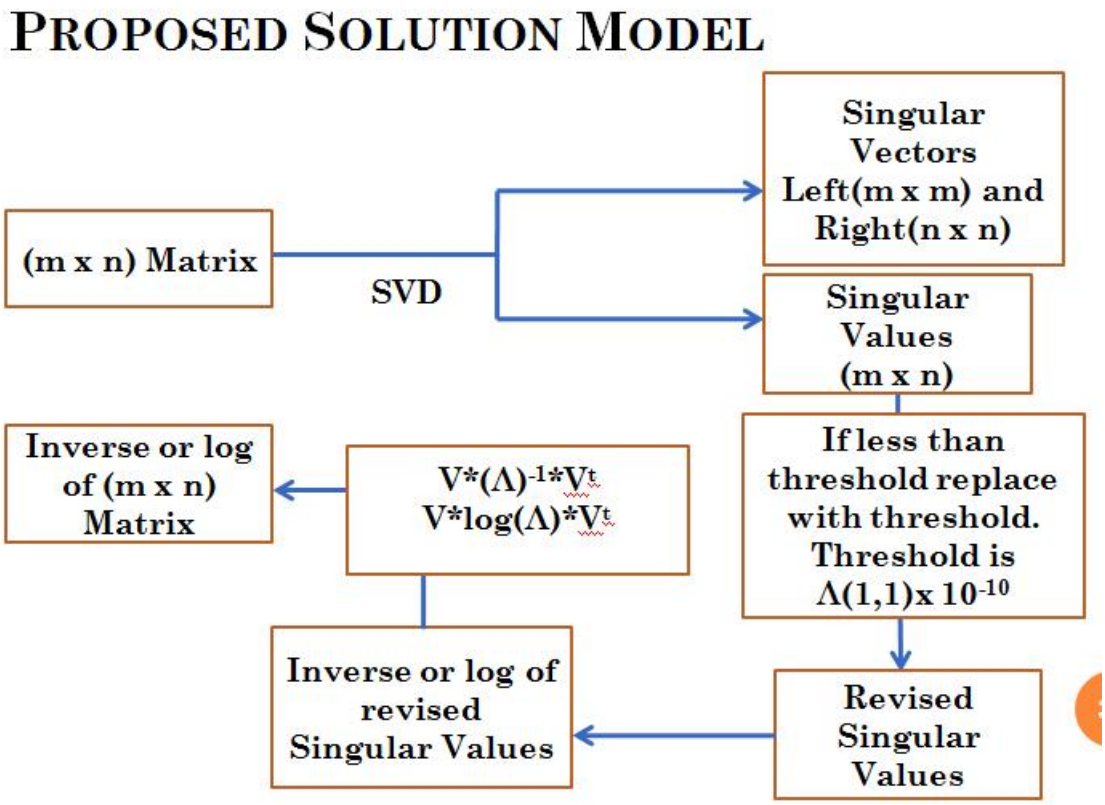

Figure 4.6: Flow diagram of solution model for singularity problem (LDR-SVD).

the singular values of a unitary matrix by our threshold. The computational complexity of this step is O(n) linear. So from here we can say that adding the singularity problem solution to the different LDA criteria does not increase the computational complexity of the criteria. From here we can conclude that our solution is also computationally efficient.

4.8.3 Flow diagram of singularity problem solution model

In the flow diagram here the solution model to solve singularity problem is shown in (Figure 4.6).

4.9 Classification and prediction evaluation

After generating feature vectors from the three-dimensional structure files of protein complexes, we have considered them as input of the LDR-SVD program to reduce the large feature dimensions to lower feature dimensions. We have implemented the Linear Dimensionality Reduction-SVD in MATLAB.

After the reduction of features to a lower dimension we have classified different protein complexes by a quadratic Bayesian classifier and a linear Bayesian classifier. After the classification, we have calculated the prediction evaluation from the confusion matrix. We have calculated accuracy, sensitivity and specificity of our classifier. We have also generated the receiver operating characteristic (ROC) curve for our prediction model. Also to measure the effectiveness of our classification model we have calculated the Matthews correlation coefficient for our classification.

4.10 Holistic view of the methodology

Step 1: Downloading PDB files

Download the Protein Data Bank three-dimensional structure files for all the complexes for different datasets (Mintseris and Zhu) from: http://www.pdb.org/pdb/home/home.do
If it is crystal packing contact, then merge those two different models in to one PDB file

Step 2: Pre-processing of the PDB files

Pre-process the PDB files by removing all unnecessary information for our feature generation.

Keep only those lines that have information about the ATOM of the complexes.

Step 3: Calculating Solvent Accessible Surface Area

Run Naccess for each PDB files of the complex

Step 4: Calculating Interface Property features

Calculate Interface Area and Interface Area Ratio of the interface features from the NACCESS generated SASA values

Calculate Number Based Amino Acid Composition of the interface feature from the RSA file

Calculate Area Based Amino Acid Composition of the interface feature from the RSA file

Calculate Amino Acid Composition of the interface or Δ V feature by using the NOXClass formula

Step 5: Calculating Pearson's Correlation Coefficient between AA Compositions of interface and Protein surface

Calculate Number Based Amino Acid Composition of the protein surface feature from the RSA file

Calculate Area Based Amino Acid Composition of the protein surface feature from the RSA file

Calculate Amino Acid Composition of the protein surface or Δ V feature by using the NOXClass formula

Calculate Pearson's Correlation Coefficient between the previously calculated Amino Acid Composition of the interface and the Amino Acid composition of the protein surface.

Step 6: Calculating Conservation Score of the interface

Download the .grades file from Consurf-DB.

Calculate the Conservation score of each complex by summing up the conservation scores of each atom.

Weight Conservation Score of each complex by their Δ SASA value

Step 7: Feature Extraction

Apply LDR with different criteria (FDA, CDA, HDA) to reduce feature dimensions.

Step 8: Classification

Classify with a Quadratic Bayesian Classifier and a Linear Bayesian Classifier.

Calculate Accuracy, Specificity and Sensitivity for each classification.

Step 9: Prediction Evaluation

Generate ROC Curve and Matthews Correlation Coefficient to evaluate the effectiveness of the prediction.

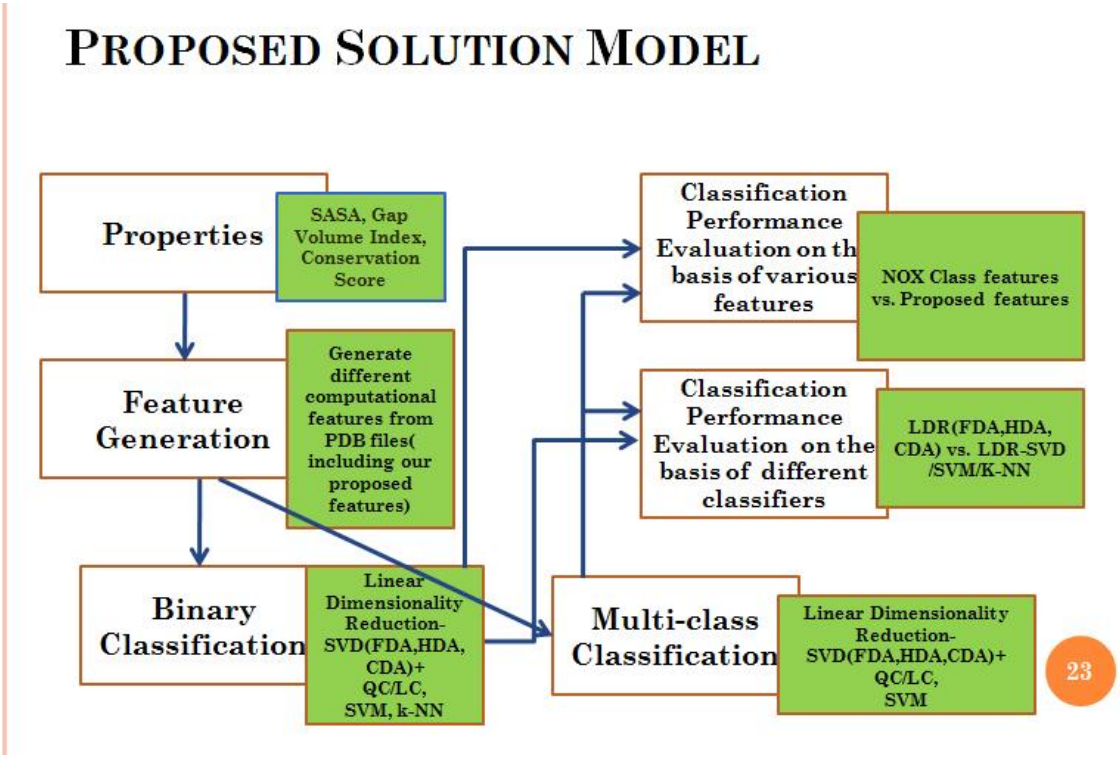

Figure 4.7: Flow diagram of the whole process of prediction of different types of protein-protein interactions.

Chapter 5

Results and Discussion

5.1 Protein-protein interaction dataset description

Protein Data Bank (PDB) (http://www.pdb.org/pdb/home/home.do) is a database of three dimensional structural data of proteins and nucleic acids. The data stored in the Protein Data Bank database is typically obtained by X-ray Crystallography and NMR Spectroscopy and submitted by biochemists. Each protein structure published in PDB receives a four character alphanumeric identifier. We have applied our classification model on two datasets. The first one is compiled by Zhu et. al. [35], and contains three different types of protein-protein interaction complexes (those are represented by four character alphanumeric identifier) namely obligate complexes, non-obligate complexes and crystal packing complexes.

The second dataset on that we have applied our classification model is compiled by Mintseris et. al. [20] and has two different types of interaction types namely obligate and non-obligate. Every protein complex here is represented by the four character alphanumeric character of their PDB name.

Table 5.1: Obligate Zhu dataset (75 complexes).

1ahj A:B	1bjn A B	1qax A B
1b34 A B	1bo1 A B	1qbi A B
1dce A B	1brm A B	1qfe A B
1efv A B	1byf A B	1qfh A B
1gux A B	1byk A B	1qor A B
1h2a L S	1c7n A B	1qu7 A B
1luc A B	1cli A B	1smt A B
1pnk A B	1cmb A B	1sox A B
1req A B	1cnz A B	1spu A B
1tco A B	1coz A B	1trk A B
2aai A B	1cp2 A B	1vlt A B
1a0f A B	1dor A B	1vok A B
1a4i A B	1f6y A B	1wgj A B
1afw A B	1gpe A B	1xik A B
1aj8 A B	1hgx A B	1xso A B
1ajs A B	1hjr A C	1ypi A B
1aom A B	1hss A B	1yve I J
1aq6 A B	1isa A B	2ae2 A B
1at3 A B	1jkm A B	2hdh A B
1b3a A B	1kpe A B	2hhm A B
1b5e A B	1msp A B	2nac A B
1b7b A C	1nse A B	2pfl A B
1b8a A B	1one A B	2utg A B
1b8j A B	1pp2 L R	3tmk A B
1b9m A B	1qae A B	4mdh A B

5.2 Experimental results

We have tested our classification model with Mintseris et. al. [20] and Zhu et. al. [35] PDB datasets. In the subsection below we first present efficiency of our proposed 40 features from experimental results for different classifiers such as SVM and LDR coupled with a Bayesian classifier. After that we show test results of LDR-SVD with datasets that have large number of features with many zeroes in them as the input feature vectors. After that, we present different prediction evaluation method that we have implemented to show our classifier's prediction effectiveness.

Table 5.2: Non-Obligate Zhu dataset (62 complexes).

1ava A C	1dow A B
1avw A B	1euv A B
1bvn T P	1i2m A B
1cse I E	1i8l A C
1eai C A	1kac A B
1f34 A B	1pdk A B
1fss A B	1qav A B
1gla F G	1tx4 A B
1kxq H A	1c0f S A
1smp I A	1zbd A B
1tab I E	1ak4 A D
1tgs I Z	1d09 A B
2ptc I E	1cqi A B
2sic I E	1fin A B
4sgb I E	1dhk A B
1agr E A	1bi7 A B
1atn A D	1wq1 R G
1b6c A B	1rrp A B
1bkd R S	1cc0 A E
1buh A B	1eg9 A B
1a4y A B	1tmq A B
1avz B C	1stf E I
1frv A B	1emv A B
3hhr A B	1uea A B
1ycs A B	1qbk B C
1cvs A C	1hlu A P
1aro L P	1itb A B
1cmx A B	1eth A B
1bml A C	1jtd A B
2pcb A B	1lfd A B
1f60 A B	1dn1 A B

Table 5.3: Crystal Packing Zhu dataset (106 complexes).

1k55	1kli	1mh9
1ual	1eyv	1ed9
1mxr	1j24	1dtd
1j98	1h1y	1ld8
1e9g	1ijy	1jlt
1iup	1exq	1ct4
1is3	1lw6	1nsz
1gy7	1m7y	1iq6
1jzl	1n3l	1i2m
1jke	1nms	1lqp
1km1	1pe0	1lqv
1ihr	1f6b	1n2e
2btc	1jp3	1i12
1eq9	1kqp	1ubk
1qf8	1j79	1g8q
1k8u	1mxi	1e87
1m7g	1my7	1jl0
1p5z	1k4i	1jr8
1e19	1jat	1qip
1k75	1f1m	1nf9
1iat	1jd0	1g60
1m9f	1nrv	1uaq
1ht9	1mvo	1ozu
1hqs	1m2d	1dmh
1b8z	1f7z	1eye
1lc5	1gyo	1i52
1gs5	1fs8	1fjj
1gve	1b67	1b16
1k20	1kzk	1e4m
1i4u	1nxm	3lyn
1k9u	1k94	1ock
1e58	1i0r	1icr
1es9	1euv	1i0d
1qkm	1ql0	1jtg
1j8b	1g2y	1elu
1kic		

Table 5.4: Obligate Mintseris dataset (115 complexes).

1b4u A:B	1fcd A:C	1kqf B:C
1dkf A:B	1eg9 A:B	3pce A:M
1fs0 E:G	1eex A:G	1gka A:B
1h2v C:Z	2kau A:C	1fxw A:F
1go3 E:F	1jkj A:B	1h8e A:D
1eex A:B	1ffu A:C	1hr6 A:B
1ccw A:B	1sgf A:B	1vcb A:B
1qdl A:B	1ffv A:B	1g8k A:B
1gpw A:B	1ldj A:B	3gtu A:B
1dtw A:B	1be3 C:A	1ytf B:D
1hsa A:B	1ezv C:F	1ktd A:B
1aui A:B	1ezv D:H	1jmz A:B
1qla A:B	2mta H:L	2kau B:C
1qlb B:C	1tbg A:E	1e9z A:B
1fm0 D:E	1h2r L:S	1qgw A:C
1mro B:C	1jb0 A:C	1li1 A:C
1mro A:C	1jb0 C:E	2ahj A:B
1e6v A:B	1jb0 C:D	1jro A:B
1ld8 A:B	1hfe L:S	1spp A:B
1l7v A:C	1k8k A:B	1vkx A:B
1dii A:C	1jwh A:C	1k8k C:G
1l9j C:H	1a6d A:B	1hxm A:B
1jmx A:G	1ir1 A:S	1req A:B
1k8k A:E	4rub A:T	1kfu L:S
1dm0 A:B	1lti A:D	1jb0 A:D
1jv2 A:B	1luc A:B	1jk0 A:B
1e50 A:B	1hzz A:C	2min A:B
1k3u A:B	1dce A:B	1raf A:B
1k28 A:D	1efv A:B	1b8m A:B
1h4i A:B	1ihf A:B	1e8o A:B
1ep3 A:B	1jb0 A:E	1jb7 A:B
1c3o A:B	1b7y A:B	1prc C:H
1k8k B:F	1dj7 A:B	1hcn A:B
1k8k D:F	1jnr A:B	1poi A:B
1k8k C:F	1kqf A:B	1f3u A:B
1jk8 A:B	1h32 A:B	1cpc A:B
1mro A:B	1mjg A:M	1dxt A:B
1m2v A:B	1n98 A:B	1exb A:E
1nbw A:B		

Table 5.5: NonObligate Mintseris dataset (211 complexes).

1wq1 G:R	1d5x A:C	1bi8 A:B
1icf A:I	1i4e A:B	1buh A:B
1fle E:I	1ib1 A:E	1qo0 A:D
2prg B:C	1kyo O:W	1ugh E:I
1h59 A:B	2mta A:C	1df9 B:C
1fbv A:C	2pcc A:B	1jiw I:P
1c4z A:D	1f3v A:B	1f93 A:E
1tmq A:B	1eja A:B	1noc A:B
1bkd R:S	1lpb A:B	1es7 A:B
1evt A:C	1dtd A:B	1k5d A:C
1dn1 A:B	1eer A:B	1hwg A:B
1xdt R:T	1ibr A:B	1fg9 A:C
1t7p A:B	1i7w A:B	1ebp A:C
1zbd A:B	1f60 A:B	1du3 A:D
1go4 A:G	1itb A:B	1cmx A:B
1fbi H:X	1ay7 A:B	1euv A:B
1ar1 A:C	1dx5 A:I	1he1 A:C
1qfw A:I	1kkl A:H	1kcg A:C
1i85 B:D	4sgb E:I	1efx A:D
1osp H:O	1dev A:B	1de4 C:A
1fsk A:B	1l0o A:C	1m10 A:B
1kxt A:B	1smf E:I	1ghq A:B
1bqh A:G	1a2k A:C	1flt V:X
1jma A:B	1dfj E:I	1gxd A:C
1kac A:B	1avg H:I	1kzy A:C
1hez A:E	1k90 A:D	1ycs A:B
1sbb A:B	1g4y B:R	1gla F:G
1e6j H:P	1jch A:B	1cxz A:B
1wej F:H	1ebd A:C	2sic E:I
1bgx H:T	1e6e A:B	1jsu A:C
1fns A:H	1gaq A:B	1lb1 A:B
1ahw A:C	1f80 A:E	1is8 A:K
2jel H:P	1buv M:T	1doa A:B
1akj A:D	4htc H:I	2mta A:H
1qfu A:H	1stf E:I	1b9y A:C
2hmi A:C	2tec E:I	1gp2 A:B
1i9r A:H	1acb E:I	1g0y I:R
1a14 H:N	1e96 A:B	1ijk A:B
1bzq A:L	1qav A:B	1i4d A:D
1nsn H:S	1f02 I:T	1k5d A:B
1lk3 A:H	1tab E:I	1n2c A:E
1ezv E:X	2ptc E:I	1mah A:F
1iqd A:C	1gl1 A:I	1gcq B:C
1dee C:G	1ezx A:C	1www V:X
1ao7 A:D	1toc A:R	1i2m A:B
1gc1 C:G	1cgi E:I	1kgy A:E
1bj1 H:V	1eai A:C	1c1y A:B
1k4c A:C	1avx A:B	1gl4 A:B
1f51 A:E	1azz A:C	1d2z A:B
1bdj A:B	1bml A:C	3ygs C:P
1eay A:C	2btc E:I	1grn A:B
1kmi Y:Z	1gh6 A:B	1cs4 A:C
7cei A:B	1iod A:G	1ki1 A:B
1clv A:I	1agr A:E	1efu A:B
1ava A:C	1avz B:C	3sgb E:I
1dhk A:B	1rlb A:E	1fqv A:B
1bvn P:T	1aro L:P	1k3z A:D
1i1a A:C	1awc A:B	1m4u A:L
1l6x A:B	1fqj A:C	1m2o A:B
1qkz A:H	1ak4 A:D	1mbu A:C
1iis A:C	1i3o A:E	1fc2 C:D
1f34 A:B	1kxp A:D	1ml0 A:D
1dpj A:B	1d4x A:G	1gvn A:B
1im3 A:D	2btf A:P	1o6s A:B
1g73 A:C	1hx1 A:B	1h2k A:S
1f83 A:B	1atn A:D	1m1e A:B
1fak H:T	1dkg A:D	1o94 A:C
1jw9 B:D	1fq1 A:B	1nf5 A:B
1jtg A:B	1fin A:B	1gzs A:B
1jtd A:B	1b6c A:B	1nbf A:D

5.2.1 Comparison between NOXClass features and the proposed features

As we have explained in the methodology, we have proposed 40 new interface property features. As shown in the Table:5.6 it increased classification accuracy for both Zhu [35] and Mintseris [20] dataset.

In the NOXClass [35] paper, Zhu et. al. predicted obligate and non-obligate complexes with 75.2 % accuracy for 6 features with a SVM classifier. By classifying them with our Fisher's LDA coupled with a Quadratic Bayesian classifier we have obtained an accuracy of 78.27 % accuracy (Table:5.6). Adding the newly proposed 40 features, our CDA with a linear classifier achieved 81.83 % accuracy (Table:5.6). Thus, we can see for Zhu dataset adding our 40 newly proposed features have increased the accuracy by 6.63 %. The SVM classifier was applied with a radial basis kernel and optimized for the best values of C and Gamma. In the Table:5.6 6 features means original features that were proposed by Zhu et. al. and 26 features means after adding number based amino acid composition features. The slight decrease in accuracy is reported while adding number based amino acid composition features; this is expected because these 20 features are very small integer numbers that lower the column rank of the feature vector. Because by adding these features increase the dependency of columns among each other, this phenomenon results in slightly lower classification accuracy. In the table, 46 features means adding area based amino acid composition features. The highest accuracy achieved for each number of features are underlined in the accuracy Table :5.6.

In [20], Mintseris et al. predicted their compiled dataset with 75% accuracy with desolvation energy features. We have used their compiled dataset that we are referring to as Mintseris dataset. We have generated interface properties features from their dataset. In

Table 5.6: Comparison between NOXClass features and newly proposed features for Zhu dataset Obligate-NonObligate.

		Quadratic			Linear		
NumberofFeatures	SVM	FDA	HDA	CDA	FDA	HDA	CDA
6	75.2	78.27	70.85	70.85	78.27	68.04	68.04
26	75.06	77.88	76.03	76.03	74.80	74.41	78.09
46	79.68	63.08	72.95	72.95	63.08	81.83	81.83

this Mintseris dataset there were some protein complexes that have multiple chains such as 1qfw AB:IM. But for our study we have taken only the first chains such as 1qfw A:I and discarded the other chains. We have generated different number of features. In the Table:5.7 6 features means NOXClass [35] proposed features. Then we have added 20 number based amino acid composition of the interface features that resulted in 26 features. After this we have added area based amino acid composition of the interface features that resulted in 46 features. In the Table:5.7 it is shown that for 6 features FDA with a Quadratic classifier achieved the highest accuracy 77.96%, with 26 features CDA with linear classifier achieved 77.54 % accuracy. The slight decrease in accuracy is reported while adding number based amino acid composition features. This is expected because these 20 features are very small integer numbers that lower the column rank of the feature vector because by adding these features increase the dependency of columns among each other. This phenomenon results in slightly lower classification accuracy. For 46 features HDA and CDA with a Quadratic classifier achieved the highest accuracy of 79.25%. We have seen 1.29% increase in accuracy from 6 features to 46 features. From these results it can be concluded that our proposed 40 features have helped to predict obligate and nonobligate protein-protein interactions with higher accuracy.

We have predicted biological and crystal packing interactions with upto 92.61 % accuracy with FDA coupled with a Quadratic Bayesian classifier. We have shown the result in

Table 5.7: Comparison between NOXClass features and newly proposed features for Mintseris dataset Obligate-NonObligate.

		Quadratic			Linear		
NumberofFeatures	SVM	FDA	HDA	CDA	FDA	HDA	CDA
4	76.43	77.96	77.32	76.35	77.29	76.01	74.74
24	76.32	77.25	77.25	76.91	75.95	77.22	77.54
44	78.34	74.37	79.25	79.25	75.02	76.01	76.97

Table 5.8: Comparison between NOXClass features and newly proposed features for Zhu dataset Biological-Crystal Packing.

		Quadratic			Linear		
Number of Features	SVM	FDA	HDA	CDA	FDA	HDA	CDA
4	90.9	91.83	84.82	87.16	91.83	89.49	89.49
24	90.68	91.44	87.94	87.94	91.05	91.05	91.05
44	91.87	92.61	90.27	90.27	92.22	92.22	92.22

the table Table:5.8 are for Zhu [35] dataset's pre-classified complexes with different number of features. We have shown that the classification accuracy increases from 90.9 % to 92.61 % from the NOXClass [35] reported results. We have shown that the prediction accuracy increases from 4 features(the best is 91.83%) to 44 features upto 92.61 %. Thus, from the results in the Table:5.8 we conclude that our proposed 40 features (number based amino acid composition and area based amino acid composition) have contributed to a better prediction of biological and crystal packing interactions. These features are proved to be efficient to predict not only obligate-nonobligate interactions but also biological-crystal packing interactions. We have predicted with different classifiers such as SVM with a Radial Basis Function Kernel optimized for C and Gamma and also with LDR coupled with Bayesian classifiers. The best accuracy 92.61 % Table:5.8 have been achieved is with FDA coupled with a Quadratic Bayesian Classifier.

5.2.2 Linear dimensionality reduction-SVD large datasets

As we have explained in methodologies, we have solved the LDR's singulairty matrix generation problem. If we used large dataset that contains many zeros as an input feature vector for LDR, it used to produce IEEE NAN values. Large dataset with many zero values would have lowered the column rank of the matrix, because of their column dependability. To avoid the problem, we have taken singular value decomposition of all the matrices that we have to take inverse of and replace their singular values with our calculated threshold values. Then, we have taken the pseudoinverse of the matrices and solved the singulairty matrix generation problem of linear disciminant analysis. To prove our solution we have tested with large datasets, the results, as reported in [19] are shown in Table:5.9. In this table, we have desolvation energies features for different dataset Mintseris [20] and Zhu [35] datasets and they contain up to 210 columns as the input feature vector. More details about these features can be found in [19]. During the execution of the program there was no singularity matrix generation problem arises. Also in the reduced feature set we have not obtained any complex numbers. From this experiment Table:5.9 we can conclude that LDR-SVD have solved the issue of singularity matrix generation problem in different LDA criteria and their subsequent effects such as complex number generation problem in the reduced feature set. In the Table:5.9 the column marked with "Name" describes the name of the dataset, column that marked with "NOF" declares Number of features in the dataset, and the column marked with "S.Problem" declares that if any singularity problem happens with the dataset or not during the execution. All these input feature vectors are reduced to dimension 20 prior to classification.

Table 5.9: Linear Dimensionality Reduction-SVD with Large Datasets Stress test.

			Quadratic			Linear		
Name	NOF	S.Problem	FDA	HDA	CDA	FDA	HDA	CDA
mintAtomNosasaOao	171	NO	69.16	79.31	79.31	72.22	80.08	80.08
mintAtomSasaOao	171	NO	68.77	77.97	77.59	70.50	78.16	78.35
mintResidueNosasaOao	210	NO	68.97	72.99	75.29	70.31	78.74	78.74
mintResidueSasaOao	210	NO	69.92	77.97	77.97	68.77	72.41	72.22
zhuAtomNosasa	171	NO	51.82	57.66	80.29	54.01	50.36	74.45
zhuAtomSasa	171	NO	50.36	56.93	77.37	55.47	51.82	72.99
zhuResidueNosasa	210	NO	64.96	67.88	56.93	67.88	70.07	74.45
zhuResidueSasa	210	NO	55.47	60.58	57.66	56.93	58.39	70.07

5.2.3 Prediction evaluation different tools

As we have explained in the methodology section we have evaluated our classifier efficiency by different measurement such as specificity, sensitivity and Matthews correlation coefficient. In the specificity table Table:5.10 it is shown that for 24 features and for FDA and CDA criteria coupled with Linear Classifier have achieved the highest specificity of 97.48 %. In the sensitivity Table:5.11 it is shown that for 44 features HDA and CDA criteria coupled with linear classifier achieved the highest sensitivity for the classifier predicting biological and crystal packing interactions. In the Matthews correlation coefficient Table:5.12 it is shown that for 44 features FDA criteria coupled with a Quadratic Bayesian classifier have achieved the highest Matthews correlation coefficient. We also presented the ROC curve for different number of features and for CDA criteria coupled with a Quadratic Bayesian classifier.

Table 5.10: Comparison between NOXClass features and newly proposed features for Zhu dataset Biological-Crystal Packing Specificity.

	Quadratic			Linear		
Number of Features	FDA	HDA	CDA	FDA	HDA	CDA
4	95.31	93.40	91.94	95.31	94.35	94.35
24	96.00	93.40	91.41	97.48	96.72	97.48
44	94.70	93.70	93.91	94.66	93.33	93.33

Table 5.11: Comparison between NOXClass features and newly proposed features for Zhu dataset Biological-Crystal Packing Sensitivity.

	Quadratic			Linear		
NumberofFeatures	FDA	HDA	CDA	FDA	HDA	CDA
4	88.37	79.14	82.71	88.37	84.96	84.96
24	87.12	90.48	90.48	85.93	85.93	85.93
44	90.40	88.33	88.00	89.68	90.98	90.98

Table 5.12: Comparison between NOXClass features and newly proposed features for Zhu dataset Biological-Crystal Packing Matthews Correlational Coefficient.

	Quadratic			Linear		
NumberofFeatures	FDA	HDA	CDA	FDA	HDA	CDA
4	0.08	0.07	0.07	0.08	0.08	0.08
24	0.08	0.08	0.08	0.08	0.08	0.08
44	0.09	0.08	0.08	0.08	0.08	0.08

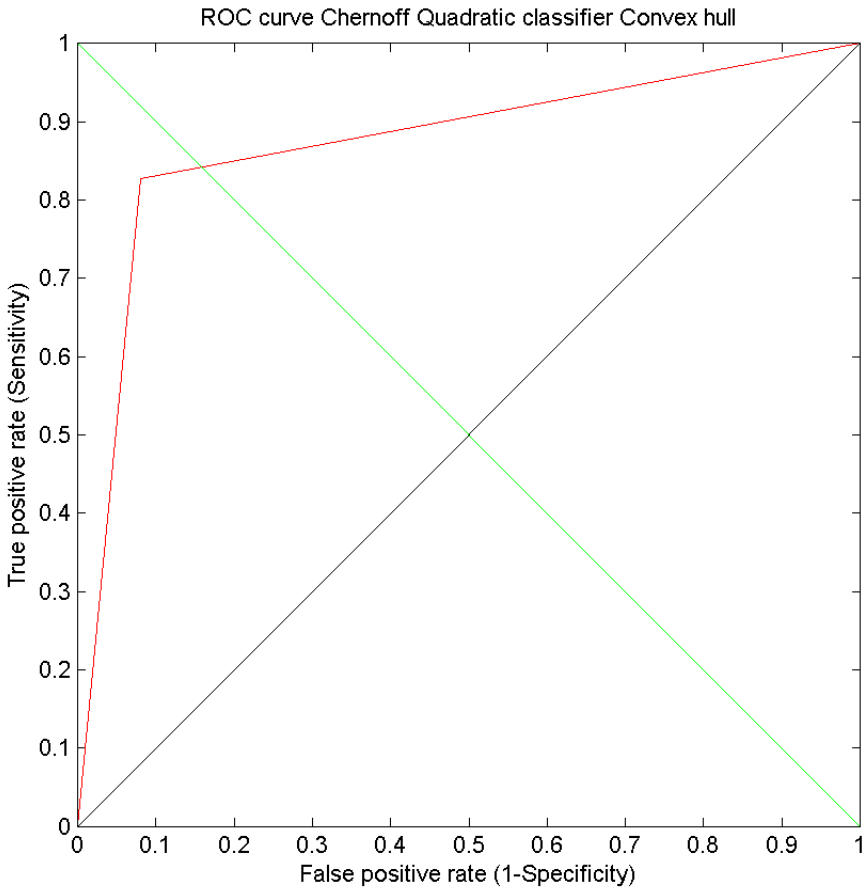

Figure 5.1: Receiver Operating Characteristic Curve for biological-nonbiological classification with CDA quadratic classifier for 4 features.

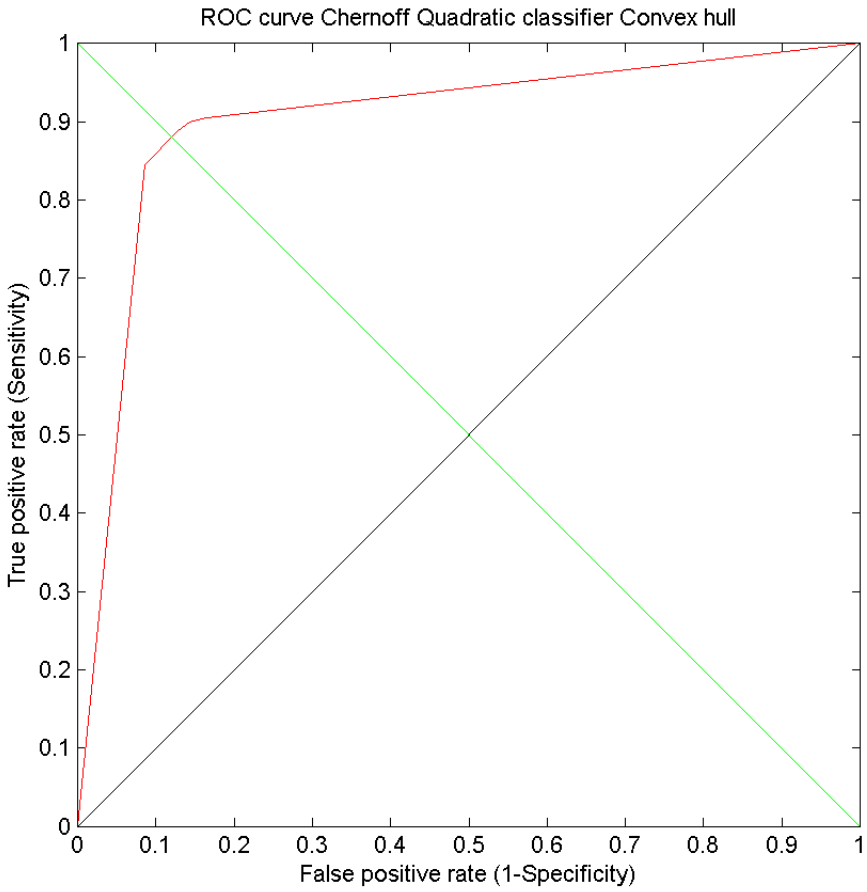

Figure 5.2: Receiver Operating Characteristic Curve for biological-nonbiological classification with CDA quadratic classifier for 24 features.

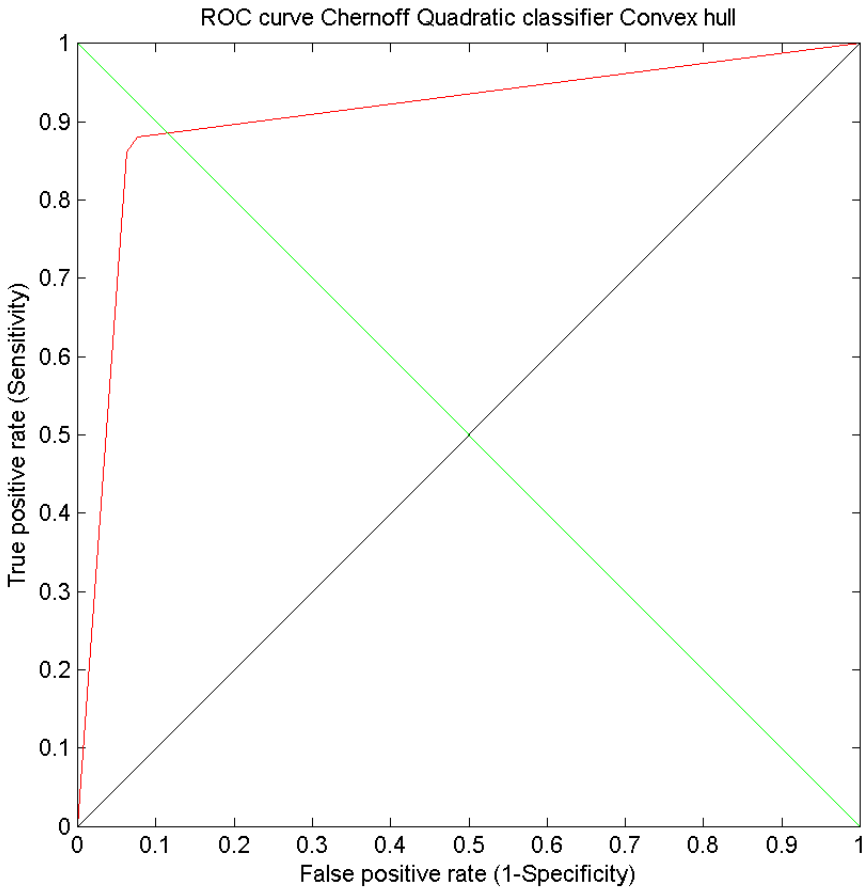

Figure 5.3: Receiver Operating Characteristic Curve for biological-nonbiological classification with CDA quadratic classifier for 44 features.

Chapter 6

Conclusions and future works

6.1 Summary of contributions

In this thesis, we have investigated interface properties to predict different types of protein-protein interactions. We have proposed 40 new interface property features of protein-protein interactions. The first 20 proposed features are number based amino acid compositions and the next 20 features are area based amino acid composition. We have shown an increased prediction accuracy for predicting obligate and nonobligate interactions for Zhu's dataset up to 81.83 %, that is, a 6.63 % increase in accuracy from NOXClass's reported result. From here it can be concluded that our newly proposed 40 interface property features are efficient to predict obligate and non obligate interactions. We have also shown an increase in prediction accuracy for obligate and non obligate interactions for Mintseris's dataset. With the newly proposed 40 features HDA criteria with a Quadratic classifier predicted obligate and non-obligate interactions with 79.25 % classification accuracy, which is higher than that of 4 features (77.96 %).

We have also shown increased prediction accuracy for biological and crystal packing

interactions. For Zhu's dataset for 46 features, our classifier have predicted biological and crystal packing interactions with 92.61 % accuracy that is an increase from NOXClass's 90.9 % accuracy. We have used multi-stage LDR and multi-stage SVM to solve this 3 class classification problem, on the first step we predicted biological and crystal packing complexes and in the second step we have predicted Obligate and Non-Obligate interaction. The results on two datasets of Zhu and Mintseris of pre classified obligate and nonobligate complexes show that the LDR schemes coupled with a Quadratic Bayesian classifier achieves the best overall classification performance, even better than an SVM kernel with C and Gamma optimized. The result on Zhu's dataset for pre classified biological and crystal packing complexes show that the FDA coupled with quadratic Bayesian classifier predicts biological and nonbiological interactions with the best accuracy.

In this thesis we have also solved singularity matrix problem of LDR. We have proposed a solution that performs singular value decomposition of each matrix that we have to take the inverse of within different LDA criteria. This scheme solves the problem of singularity matrix generation for heteroscedastic discriminant analysis and Chernoff discriminant analysis and do not produce any complex number in the reduced feature set. We have also implemented some prediction evaluation tool such as receiver operating characteristic curve, Matthews' correlation coefficient to evaluate and visualize the effectiveness of our classification model.

6.2 Future work

The prediction approach that is discussed in the thesis can be applied to predict other types of protein-protein interaction classification mechanism. Another problem that deserves investigation is to devise a strategy to discriminate more than two classes with multi-class

LDR. Other interesting problems that deserve investigation are the use of this approach in different protein-protein interaction classification problems including intra and inter domains, homo and hetero oligomers, and the use of other features such as geometric features such as shape of the structure of the interface, planarity and roughness, and other statistical and physiochemical properties such as residue and atom vicinity, secondary structure elements and domains, hydrophobicity, salt bridges are among others.

Bibliography

[1] M. A. and B. R. Protein interaction methods toward an endgame. *Science*, 284(5422): 1948–1950, 1999.

[2] T. A. The Protein structure project. *The Rigaku Journal*, 16(5422):1948–1950, 1999.

[3] A. Armon, D. Graur, and Ben-TalN. Consurf:an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J Mol Biol*, 307(447–463), 2001.

[4] A. B., A. T., L. B., R. S., and E. K. Quantification of protein half-lives in the budding yeast proteome. *PNAS*, 103(35):13004–13009, 2006.

[5] R. Duda and P. Hart. *Pattern Classification and Scene Analysis*. John Wiley and Sons, Inc., 1973.

[6] P. H., H. K., and T. J. Discriminating between homodimeric and monomeric proteins in the crystalline state. *Proteins*, 41(9):47–57, 2000.

[7] S. Hubbard and J. Thornton. Naccess, 1993. URL `http://www.bioinf.manchester.ac.uk/naccess/` .

[8] N. I. and T. J. Structural characterisation and functional significance of transient protein-protein interactions. *J Mol Bio*, 325(5):991–1018, 2003.

[9] J. J. Specific versus non-specific contacts in protein crystals. *Nat Struct Biol*, 4(12): 973–974, 1997.

[10] J. J. and R. F. Protein-protein interaction at Crystal Packing Contacts. *Protein*, 23(4): 580–587, 1995.

[11] J. Janin. *Kinetics and thermodynamics of protein-protein interactions from a structural perpective*. Oxford University Press, 2000. Protein-Protein Recognition, pp. 344.

[12] S. Jones and J. M. Thornton. Principles of protein-protein interactions. *Proc. Natl Acad. Sci, USA*, 93(1):13–20, 1996.

[13] S. Jones and J. M. Thornton. *Protein-Protein Recognition*, chapter Analysis and classification of protein-protein interactions from a structural perspective. Oxford University Press, 2000.

[14] G. K., T. C., and N. R. Analysis of ordered and disordered protein complexes reveals structural features discriminating between stable and unstable consumers. *J Mol Bio*, 341(5):1327–1341, 2004.

[15] H. K. and T. J. PQS:a protein quaternary structure file server. *Trends Biochem Sci*, 23 (9):358–361, 1998.

[16] L. C. L., C. C., and J. J. The atomic structure of protein-protein recognition sites. *J Mol Bio*, 285(5):2177–2198, 1999.

[17] R. Laskowski. Surfnet: a program for visualizing molecular surfaces , cavities and intermolecular interactions. *J Mol Graph*, 3(5):323:30:307–8, 1995.

[18] M. Loog and P. Duin. Linear Dimensionality Reduction via a Heteroscedastic Extension of LDA: The Chernoff Criterion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(6):732–739, 2004.

[19] A. M., M. M., R. L., R. M., and B. S. Prediction of biological protein-protein interaction using atom type and amino acid properties. *Wiley VCH Proteomics*, 2011(11): 1–10, 2011.

[20] J. Mintseris and Z. Weng. Structure, function, and evolution of transient and obligate protein-protein interactions. *Proc Natl Acad Sci, USA*, 102(31):10930–10935, 2005.

[21] J. Mintseris and Z. Weng. Structure, Function, and Evolution of Transient and Obligate Protein-protein Interactions. *Proceedings of the National Academy of Sciences, USA*, 102(31):10930–10935, 2005.

[22] I. Nooren and J. Thornton. Diversity of protein-protein interactions. *EMBO Journal*, 22(14):3846–3892, 2003.

[23] C. O. and A. P. Protein-protein crystal Packing Contacts. *Protein Sci*, 6(10):2261–2263, 1997.

[24] B. R., C. P., R. F., and J. J. A dissection of specific and non-specific protein-protein interfaces. *J Mol Bio*, 336(4):943–955, 2001.

[25] R. R., A. F., and A. P. A structural perspective on protein-protein interactions. *Curr Open Strcut Biol*, 14(3):313–324, 2004.

[26] L. Rueda and M. Herrera. Linear Dimensionality Reduction by Maximizing the Chernoff Distance in the Transformed Space. *Pattern Recognition*, 41(10):3138–3152, 2008.

[27] R. B. Russell, F. Alber, P. Aloy, F. P. Davis, D. Korkin, M. Pichaud, M. Topf, and A. Sali. A structural perspective on protein-protein interactions. *Curr Opin Struct Biol*, 14(3):313–324, 2004.

[28] D. S., I. G., B. S., L. C., and B. J. Extent and nature of contacts between protein molecules in crystal lattices and between subunits of protein oligomers. *Proteins*, 28 (4):494–514, 1997.

[29] D. S., K. O., S. N., and R. N. Interaction preferences across protein-protein interfaces of obligatory and non-obligatory components are different. *BMC Struct Biol*, 15(5): 1327–1341, 2005.

[30] J. S. and T. J. Analysis of protein-protein interaction sites using surface patches. *J Mol Bio*, 272(4):121–132, 1997.

[31] J. S. and T. J. Prediction of protein-protein interaction sites using patch analysis. *J Mol Bio*, 272(1):133–143, 1997.

[32] S. Theodoridis and K. Koutroumbas. *Pattern Recognition*. Elsevier Academic Press, third edition, 2006.

[33] V. W. and T. J. Conservation helps to identify biologically relevant crystal contacts. *J Mol Bio*, 313(2):399–416, 2001.

[34] O. Y. and R. B. Analysing six types of protein-protein interfaces. *J Mol Bio*, 325(2): 377–87, 2003.

[35] H. Zhu, F. Domingues, I. Sommer, and T. Lengauer. Noxclass: Prediction of protein-protein interaction types. *BMC Bioinformatics*, 7(27), 2006. doi:10.1186/1471-2105-7-27.

Vita Auctoris

Sridip Banerjee was born in 1987 in Calcutta, India. He received his Bachelors degree in Information Technology from West Bengal University of Technology, Calcutta, India in 2009. His research interests include Pattern Recognition, Protein-Protein Interaction, Data Mining, Machine Learning and Bio-informatics.